Exhibit 2

PLAINTIFFS' OPPOSITION TO DEFENDANTS' MOTION TO EXCLUDE PLAINTIFFS' EXPERTS' GENERAL CAUSATION OPINIONS FOR FAILURE TO ACCOUNT FOR SECTION 230 AND THE FIRST AMENDMENT

Highly Confidential (Competitor)

SUPERIOR COURT OF THE STATE OF CALIFORNIA
FOR THE COUNTY OF LOS ANGELES

COORDINATION PROCEEDING SPECIAL TITLE [RULE 3.400] SOCIAL MEDIA CASES	JUDICIAL COUNCIL COORDINATED PROCEEDING NO. 5255 Judge: Hon. Carolyn B. Kuhl **EXPERT REPORT OF DR. EMILIO FERRARA**
THIS DOCUMENT RELATES TO: ALL CASES	

TABLE OF CONTENTS

I. Introduction and Summary of Qualifications

1. My name is Emilio Ferrara. I am a Professor of Computer Science at the University of Southern California, Associate Chair at the Thomas Lord Department of Computer Science, and the inaugural (ad interim) Director of the Interdisciplinary Data Science Program at the USC School of Advanced Computing. I have joint appointments in Communication (USC Annenberg) and Preventive Medicine (USC Keck School of Medicine). I am also Principal Scientist at the Information Sciences Institute, Principal Investigator at the USC-ISI Machine Intelligence and Data Science (MINDS) center, and Director of the PhD program in Computer Science.

2. I have summarized in this section my educational background, career history, publications, and other relevant qualifications. A copy of my CV containing additional details relating to my background and qualifications is attached as Exhibit A to this report.

A. Educational Background

3. I attended the University of Messina (Italy), where I received a bachelor's degree in Computer Science in 2006, and a Master's degree in Computer Science in 2008. I completed my doctoral studies by the end of 2011, which focused on Machine Learning, with a concentration on the study of online social networks. My doctoral thesis titled "Mining and analysis of online social networks" was successfully defended in March 2012. My PhD degree was awarded the special honorific title of Doctor Europaeus, which indicates that the PhD student has met specific requirements of the European University Association related to international collaboration, mobility, and multilingualism.[1] During my graduate studies, I published 13 peer-reviewed research papers. I have since published over 200 additional peer-reviewed research papers.

B. Professional Experience

4. In 2010, I served as a visiting scholar at Technische Universität Wien in Vienna, Austria, where I was involved with research focusing on Data Mining and Machine Learning.

[1] *See Doctor Europaeus*. (n.d). Università degli Studi di Udine. https://www.uniud.it/en/research/do-research/doctorate-res/international-programmes/doctor-europaeus#:~:text=The%20Doctor%20Europaeus%20certification%20is,qualification%20issued%20by%20international%20institutions.

While there, I worked as an intern at Lixto Software GmbH (later acquired by McKinsey), where I was involved with research to develop Web crawler bots for large-scale, distributed data collection from online services. This work appeared in 3 peer-reviewed research papers.

5. From 2011-2012, I was a visiting scholar at Royal Holloway, University of London in Egham, England, where I was involved with research focusing on Network Science, Computational Biology, and Machine Learning. During this period, I developed techniques for the clustering of complex network data and applied these methodologies to social and biological networks. This research has appeared in over a dozen peer-reviewed research articles. I implemented these techniques and released them as open-source software, now used by hundreds of research papers spanning applications in science, engineering, and business.

6. At Indiana University Bloomington (2012-2015), I conducted research focusing on Network Science, Data Science, Data Mining, and Machine Learning. My roles included postdoctoral research fellow, assistant research scientist, and research assistant professor within the School of Engineering and Computing and the Indiana University Networks Institute (IUNI). During this period, I published approximately 40 peer-reviewed research papers, which laid the foundation for the study of Twitter's content moderation systems. My focus was the detection of manipulation and abuse on the Twitter service. I contributed to the ideation and design, as a co-inventor, of the "Bot Or Not" (later renamed "Botometer") framework for the detection of bots on Twitter. I used this and other machine learning techniques I invented to study a multitude of social phenomena on Twitter, including but not limited to (i) the coordination of Twitter users revolving around the "Occupy Wall Street" social movements of 2012; (ii) the dynamics of evolution of Twitter trends; (iii) the behavioral evolution of Twitter users revolving around the "Gezi Park" protests of 2013 in Turkey; (iv) the issue of misogynistic language and toxic abuse of Twitter users; (v) the issue of emotional contagion and the dynamic of spread of negative sentiment on Twitter; (vi) the issue of manipulation and abuse, specifically the manufacturing of alarming misinformation, on Twitter in the context of the 2014 Ebola public health crisis.

7. In 2015, I began working at the University of Southern California as a computer

scientist, conducting research into Network Science, Data Science, Data Mining, and Machine Learning. In 2016, I transitioned to the role of research assistant professor at USC, and in 2020, I became a tenured associate professor. In 2022, I was promoted to full professor, the same role I have today. I currently hold appointments as professor of computer science, communication, and public health, as well as associate chair of the Department of Computer Science.

8. In 2021, I also began working with Amazon as a visiting academic in their Alexa AI department. In this role, I drive research initiatives to prevent Alexa from providing inappropriate responses to customers. I do so by collaborating with Amazon's Alexa team in charting the taxonomy of types of problematic content (e.g., hate speech, spam and scams, health misinformation, adult content, violent content, racially-charged or other types of biased content, and more) that should not be digested by, consumed through, or exposed to Alexa AI. I also design machine learning techniques to (i) prevent indexing problematic content from the Web and social media services; (ii) detect and filter out problematic content early on from the indexed data; (iii) detect questions that are unsuitable to a machine-generated answer; (iv) detect unsuitable answers to specific audiences (e.g., children); (v) detect social media trends that might be reflected in inquiries to Alexa; (vi) anticipate inappropriate content from social media feeds.

9. Since 2015, I have developed and used a plethora of machine learning and AI techniques to study a multitude of problems on social media services, including but not limited to (i) detecting online extremism and radical propaganda by the ISIS terrorist group on social media; (ii) detecting spam in electronic-cigarette advertising campaigns on social media; (iii) unveiling manipulation of social media discourse revolving around the 2016 US Presidential Election; (iv) automatically detecting bots by means of an ensemble of classifiers, known as "Bot Or Not" and later "Botometer" framework; (v) participating in the Defense Advanced Research Projects Agency social media bot challenge and ranking in the top 3 teams for accuracy of bot detection; (vi) characterizing the interactions between human and bot accounts on social media; (vii) characterizing the "Macron Leaks" social media disinformation campaign surrounding the 2017 French Presidential election; (viii) characterizing information contagion manipulation by means

of social media bots; (ix) detecting early warnings of cyberattacks and cyberthreats on social media; (x) unveiling the use of bots to bolster negative and inflammatory content in the context of the 2017 Catalan referendum; (xi) analyzing the digital traces of political manipulation on social media by the Russian Internet Research Agency (IRA); (xii) developing deep learning based techniques for bot detection; (xiii) measuring spam and the effect of bots on information diffusion on social media; (xiv) unveiling the perils of manipulation of social media discourse surrounding the 2018 US Midterm elections; (xv) characterizing the behavior of Russian IRA trolls operated social media influence campaigns; (xvi) charting the history of digital spam on social media services; (xvii) characterizing linguistic cues to deception used by spammers and trolls on social media; (xviii) comparative analysis of bot partisan behavior on social media; (ixx) characterizing the evolution of bot behavior over election-related events; (xx) charting the landscape of social media cryptocurrency manipulation; (xxi) developing a reinforcement learning framework to detect spam behavior on social media; (xxii) developing a temporal behavioral dynamics model to detect bot behavior on social media; (xxiii) tracking social media related COVID-19 conversation since the inception of the pandemic to track potential manipulation and abuse; (xxiv) characterizing the implications of political polarization surrounding COVID-19 in the US-based social media discussion; (xxv) building a multimodal dataset for information credibility studies of social media and news sources; (xxvi) characterizing social media manipulation by bots and influence campaigns in the context of the 2020 US Presidential Election; (xxvii) unveiling COVID-19 misinformation influencing election-related discourse on social media; (xxviii) characterizing COVID-19 vaccine hesitancy and anti-vaccine campaigns spreading on social media; (xxix) charting the effects of echo chambers in the context of COVID-19 on social media; (xxx) identifying coordinated accounts on social media via hidden influence modeling; (xxxi) characterizing online engagement with election-related disinformation and conspiracies on social media; (xxxii) auditing the effects of algorithmic bias in news-feed ranking on social media; (xxxiii) constructing a large-scale annotated dataset of social media misinformation campaigns; (xxxiv) comparative analysis of bots and humans during the COVID-19 pandemic on social media;

(xxxv) surveying influence campaigns during the Stop Asian Hate Movement on social media; (xxxvi) tracking the discourse and influence campaigns on the war between Ukraine and Russia on social media; (xxxvii) charting the landscape of information manipulation in the context of the Israel Hamas conflict; (xxxviii) inferring political leaning of social media users from linguistics and psychographic cues; (xxxix) modeling of user attraction to politically extreme content on social media; (xl) understanding the emotional variance of geolocated social media posts; (xli) studying the effect of user exposure to propaganda on social media; (xlii) modeling the interconnected nature of content moderation and exposure to harmful content on social media; (xliii) unveiling the effects of content censorship by studying Chinese social media; (xliv) modeling the mechanism of adoption and motivators to produce toxic content online; (xlv) modeling the adoption of information after minimal exposure on social media; (xlvi) using AI moderators to improve the quality and limit harmful content exposure online. Since 2015, I have developed and used a plethora of machine learning and AI techniques to study a multitude of problems on social media services, including but not limited to (i) detecting online extremism and radical propaganda by the ISIS terrorist group on social media; (ii) detecting spam in electronic-cigarette advertising campaigns on social media; (iii) unveiling manipulation of social media discourse revolving around the 2016 US Presidential Election; (iv) automatically detecting bots by means of an ensemble of classifiers, known as "Bot Or Not" and later "Botometer" framework; (v) participating in the Defense Advanced Research Projects Agency social media bot challenge and ranking in the top 3 teams for accuracy of bot detection; (vi) characterizing the interactions between human and bot accounts on social media; (vii) characterizing the "Macron Leaks" social media disinformation campaign surrounding the 2017 French Presidential election; (viii) characterizing information contagion manipulation by means of social media bots; (ix) detecting early warnings of cyberattacks and cyberthreats on social media; (x) unveiling the use of bots to bolster negative and inflammatory content in the context of the 2017 Catalan referendum; (xi) analyzing the digital traces of political manipulation on social media by the Russian Internet Research Agency (IRA); (xii) developing deep learning based techniques for bot detection; (xiii)

measuring spam and the effect of bots on information diffusion on social media; (xiv) unveiling the perils of manipulation of social media discourse surrounding the 2018 US Midterm elections; (xv) characterizing the behavior of Russian IRA trolls operated social media influence campaigns; (xvi) charting the history of digital spam on social media services; (xvii) characterizing linguistic cues to deception used by spammers and trolls on social media; (xviii) comparative analysis of bot partisan behavior on social media; (ixx) characterizing the evolution of bot behavior over election-related events; (xx) charting the landscape of social media cryptocurrency manipulation; (xxi) developing a reinforcement learning framework to detect spam behavior on social media; (xxii) developing a temporal behavioral dynamics model to detect bot behavior on social media; (xxiii) tracking social media related COVID-19 conversation since the inception of the pandemic to track potential manipulation and abuse; (xxiv) characterizing the implications of political polarization surrounding COVID-19 in the US-based social media discussion; (xxv) building a multimodal dataset for information credibility studies of social media and news sources; (xxvi) characterizing social media manipulation by bots and influence campaigns in the context of the 2020 US Presidential Election; (xxvii) unveiling COVID-19 misinformation influencing election-related discourse on social media; (xxviii) characterizing COVID-19 vaccine hesitancy and anti-vaccine campaigns spreading on social media; (xxix) charting the effects of echo chambers in the context of COVID-19 on social media; (xxx) identifying coordinated accounts on social media via hidden influence modeling; (xxxi) characterizing online engagement with election-related disinformation and conspiracies on social media; (xxxii) auditing the effects of algorithmic bias in news-feed ranking on social media; (xxxiii) constructing a large-scale annotated dataset of social media misinformation campaigns; (xxxiv) comparative analysis of bots and humans during the COVID-19 pandemic on social media; (xxxv) surveying influence campaigns during the Stop Asian Hate Movement on social media; (xxxvi) tracking the discourse and influence campaigns on the war between Ukraine and Russia on social media; (xxxvii) charting the landscape of information manipulation in the context of the Israel Hamas conflict; (xxxviii) inferring political leaning of social media users from linguistics and psychographic cues; (xxxix) modeling of user attraction to

politically extreme content on social media; (xl) understanding the emotional variance of geolocated social media posts; (xli) studying the effect of user exposure to propaganda on social media; (xlii) modeling the interconnected nature of content moderation and exposure to harmful content on social media; (xliii) unveiling the effects of content censorship by studying Chinese social media; (xliv) modeling the mechanism of adoption and motivators to produce toxic content online; (xlv) modeling the adoption of information after minimal exposure on social media; (xlvi) using AI moderators to improve the quality and limit harmful content exposure online.

10. On November 7, 2016, I published the paper titled *Social bots distort the 2016 US Presidential election online discussion*. This is the only peer-reviewed research paper to appear before the 2016 US Presidential Election of November 8, 2016. The work shows early signs of foreign state-sponsored operations on social media by means of bots and spam accounts. This work was utilized in the US Senate's investigation into the 2016 election and helped expose the now well-known foreign-sponsored efforts to interfere with the online political discussion.

11. My research has informed policy and regulation at both the federal and state level. It has been referenced in laws and bills such as (i) the California "SB 1001 Bots: disclosure" that was ratified into law as the B.O.T. Act ("Bolstering Online Transparency") as of July 2019 and (ii) the "S.2125 - Bot Disclosure and Accountability Act of 2019", introduced by former Senate Judiciary Committee Ranking Member Dianne Feinstein.

12. Since 2015, I have published over 200 peer-reviewed research papers. As of June 4, 2025, my research has been cited over 25,000 times according to Google Scholar. By this measure of impact, I am one of the top 5 most cited researchers worldwide in the area of "Human-Centered AI," top 20 most cited in "Social Computing" and "Computational Social Science," and top 30 most cited in the areas of "Social Media," and top 200 most cited researchers in the broad area of "Data Science."

13. Since 2015, I have secured over $60 million in research funding, the vast majority of which is funded by the US Department of Defense (DOD), Defense Advanced Research Projects Agency, or government intelligence agencies (e.g., Intelligence Advanced Research Projects

Activity). This funding has supported research in social media manipulation, bot detection, influence campaign detection, and online service abuse. In addition, certain technologies I have developed have been transitioned, or are in the course of transitioning, into DOD operations by means of Technology Transition awards. Most prominently, the Twitter bot detection techniques I invented have been transitioned to the US government and defense agencies under the Office of Naval Research SBIR/STTR Contract, "DOISAC: Detecting Orchestrated Information and Synthetic Account Campaigns," of which I was the principal investigator.

C. Publications

14. I have authored, co-authored, and contributed to over 250 academic papers and publications, mostly in data mining and machine learning, particularly with respect to social media networks. A copy of my CV listing my publications and contributions is attached as Exhibit A to this report.

II. Plaintiffs' Allegations and Summary of Opinions

A. Plaintiffs' Allegations Regarding Content Moderation

15. Plaintiffs allege that Meta has knowingly failed to implement adequate safeguards that prevent, detect, or remove allegedly harmful content on Meta's family of apps.[2] Plaintiffs allege that such failure has led to a mental health crisis among Meta's users, especially teens.[3]

B. Nature of Assignment and Summary of Opinions

16. I have been retained by Covington & Burling LLP, counsel for Meta Platforms, Inc., to provide my expert opinions on Meta's content moderation systems, practices, and related issues on Facebook and Instagram. In particular, I have been asked to assess the reasonableness of Meta's content moderation systems in light of published academic literature and the challenges of maintaining a content moderation system on social media services at the scale a company like Meta operates. My review of Facebook and Instagram's content moderation systems is consistent

[2] Master Complaint (Personal Injury) at 46-47, JCCP No. 5255 (Cal. Super. Ct. May 15, 2023).

[3] Master Complaint (Personal Injury) *passim*, *Social Media Cases*, JCCP No. 5255 (Cal. Super. Ct. May 15, 2023).

with how I review such technologies for my research.

17. I am being compensated at my customary rate of $1,200.00 per hour. My compensation is not contingent on the outcome of this litigation.

18. I have reviewed the expert reports of Drs. Bagot, Cingel, Christakis, Goldfield, Lembke, Mojtabai, Murray, Telzer and Twenge disclosed on April 18, 2025 and May 16, 2025. These experts have, in general, made high-level allegations about the adequacy of Meta's content moderation efforts, and to the extent that they have done so, the opinions set forth in this report are responsive. I understand that in subsequent rounds of expert reports, Plaintiffs' experts (those above or potentially additional experts) may provide more specific opinions on Meta's content moderation systems, and I reserve the right to respond further to any such opinions.

19. I have also evaluated a purported stress test done by Mr. Arturo Bejar. I based my evaluation on my extensive experience in algorithmic auditing, having performed several such studies myself on social media services such as Twitter/X, TikTok, Reddit, and others.

20. In summary, I have developed the following opinions based upon my review of the relevant academic literature, filings in this case, documents and testimony produced in discovery, written discovery, and my knowledge and expertise. Attached as Exhibit B, please see a complete list of the sources I have reviewed to develop my opinions. For details of the cases in which I have testified within the last four years, please see Exhibit C.

21. Meta has implemented one of the most technically advanced and operationally mature content moderation infrastructures in the industry, in which it has invested substantial resources to develop and continually refine.

22. Meta's Community Standards Enforcement Reports ("CSERs") have provided a scientifically sound and statistically robust mechanism for evaluating the effectiveness of content moderation.

23. Meta's content moderation systems prioritize the removal of high-severity content, such as suicide & self-injury content and child safety content, and demonstrate exceptionally high proactive detection rates in these categories.

24. The company has demonstrated a transparent and iterative approach to measuring, refining, and reporting the performance of its content moderation systems.

25. Meta's content moderation systems are not static but are instead designed to evolve in response to emerging harms, adversarial tactics, cultural considerations, and regulatory constraints.

26. Allegations that Meta's recommendation systems promote harmful content are not grounded in scientifically valid methodologies.

27. Meta's content moderation systems do not and could not lead to a perfect result in which no policy-violating content is present on Meta's services—that would be, in my opinion, technically impossible given the immense challenges associated with the massive volume and varied nature of the content on the services, not to mention the efforts of some adversarial, oftentimes criminal, users to try to circumvent Meta's content moderation systems and intentionally post policy-violating content. But, in my opinion, Meta's content moderation systems are robust, reasonably designed, and continuously improving.

III. Background

A. Overview of Content Moderation on Social Media Services

28. Social media provides a service for users to create and post content and to communicate with other users. Most social media companies, including Meta, create and enforce policies to prohibit or limit certain types of content on the services. This report refers to the systems and processes used by Meta to enforce such policies as "content moderation."

29. Content moderation, in the context of social media, is multifaceted and ever changing, involving various stakeholders, including service operators, users, regulators, and society at large. It is an interplay between the technical mechanisms that underpin the services and the policy frameworks that guide their use and governance.

30. Different services define content moderation differently, which at times can make comparing policies or technological solutions across different service providers extremely

complex.[4] Irrespective of the definition, content moderation on user generated platforms is an important aspect of the user experience because it dictates the content a user can post and the content a user is likely to see.

31. In the initial era of social media services (in the early 2000s), services had relatively small user bases with a relatively limited amount of content being shared. As such, content moderation took place on a relatively small scale and often relied upon simple rule-based approaches.[5] Typically, small teams of human curators manually reviewed content,[6] a strategy that was not sustainable in the long term for growing platforms, in part because of sheer scalability challenges, and in part due to transparency and power imbalance issues, with the potential for other issues such as censorship or collusion among users.[7] Although rudimentary, these early systems established the foundations for what would become more sophisticated moderation systems.[8]

32. As user bases grew into the millions, the task of content moderation and feed curation quickly surpassed human review scalability due to the growing volume of content being shared. To address scale, services turned to automated systems as basic machine learning began to take root.[9] These early systems, however, were often simplistic and usually failed to understand

---

[4] *See* Pierri, Francesco, and Stefano Ceri. *False news on social media: a data-driven survey*. ACM Sigmod Record 48.2 (2019): 18-27; Sharma, Karishma, et al. *Combating fake news: A survey on identification and mitigation techniques*. ACM Transactions on Intelligent Systems and Technology (TIST) 10.3 (2019): 1-42.

[5] Jiang, J. A., Middler, S., Brubaker, J. R., & Fiesler, C. (2020, October*). Characterizing community guidelines on social media platforms*. In Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing (pp. 287-291).

[6] *See* Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, which documents the role of early human moderation on platforms like YouTube and Facebook.

[7] *See* Jhaver, S., Bruckman, A., & Gilbert, E. (2019). *Does transparency in moderation really matter?* CHI Conference on Human Factors in Computing Systems.

[8] Riedl, M. J. (2020). *Content moderation and volunteer participation*. In The Routledge Encyclopedia of Citizen Media (pp. 93-98). Routledge.

[9] Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 2053951719897945.

complex social nuances and context, sometimes leading to public controversies around censorship and bias.[10]

33. In the early 2010s, a surge in content creation brought new challenges to content moderation. Incidents around the abuse of social media services for nefarious purposes were on the rise (e.g., malicious targeting of certain groups by attributing propaganda to members of those groups).[11]

34. Meta invested early in automated defense systems. The Facebook Immune System[12] ("FIS"), launched by early 2011, represents Meta's first attempt at a real-time integrity infrastructure, designed to detect and mitigate a wide range of abuses, including phishing, spam, and policy-violating content, across billions of daily user actions. Also in 2011, Meta began using PhotoDNA, a technology used to detect known signals of child sexual imagery, on its services.[13] Meta was one of the first large companies to broadly implement the technology and encouraged and trained other companies on how to implement it.[14]

35. Later, social media companies—Meta chief among them—began to invest in more advanced AI systems capable of detecting nuances in language and imagery, and partnerships with fact-checkers became more common. This era marked the beginning of a more proactive approach to content moderation.

B. Challenges to Content Moderation on Social Media Services

36. As social media companies began to operate globally, their services started to face

---

[10] *See* Klonick, Kate. *The new governors: The people, rules, and processes governing online speech*. Harv. L. Rev. 131 (2017): 1598.

[11] *See* "Disinformation Nation: Social Media's Role in Promoting Extremism and Misinformation", 117th Congress (2021-2022), https://www.congress.gov/event/117th-congress/house-event/111407/text

[12] *See* Stein, T., Chen, E., & Mangla, K. (2011, April). *Facebook immune system. In Proceedings of the 4th workshop on social network systems.* https://css.csail.mit.edu/6.858/2014/readings/facebook-immune.pdf

[13] META3047MDL-034-00098929.

[14] META3047MDL-034-00098929.

significant challenges in adapting their policies and corresponding content moderation technology to shifting global priorities as to what harms are most pressing, as well as to the multijurisdictional regulatory changes that have followed. Coupled with and exacerbating socio-political challenges are the technical hurdles that services must overcome in order to respond to public and political concerns.

1. Socio-political Challenges

37. Societal norms and cultural perspectives shape what is deemed harmful, and these perspectives shift over time and across regions. This necessitates nuanced understandings of cultural contexts, which has led to the incorporation of regional expertise in content policy teams and the development of more localized moderation practices.[15] Certain content that was once considered benign can become problematic or harmful as social awareness evolves, and regulatory bodies respond in kind. Content moderation policies must constantly adapt to these shifting definitions of harm to remain relevant and effective.[16]

a. What harm is the most pressing is always evolving

38. Global events, political climates, and public discourse often dictate what perceived public harm is the most urgent.[17] Meta must maintain agile and responsive policies that dynamically adjust to these evolving concerns.

39. Academic research supports this observation, highlighting the fluid nature of harmful content definitions and the challenges they pose for content moderation policies.[18] For

[15] *See* Tworek, Heidi. *History explains why global content moderation cannot work*. (2021). https://www.brookings.edu/articles/history-explains-why-global-content-moderation-cannot-work/

[16] Scheuerman, M. K., Jiang, J. A., Fiesler, C., & Brubaker, J. R. (2021). *A framework of severity for harmful content online*. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-33.

[17] Chan, A. J., Redondo García, J. L., Silvestri, F., O'Donnell, C., & Palla, K. (2023). *Enhancing Content Moderation with Culturally-Aware Models*. arXiv e-prints, arXiv-2312.

[18] Arora, A., Nakov, P., Hardalov, M., Sarwar, S. M., Nayak, V., Dinkov, Y., ... & Augenstein, I. (2023). *Detecting harmful content on online platforms: what platforms need vs. where research efforts go*. ACM Computing Surveys, 56(3), 1-17.

instance, Dr. Amit Sheth and collaborators authored a prominent study that emphasizes that online problematic content is multi-dimensional and context-sensitive, making its detection challenging.[19] The authors argue that definitions of toxic content extend beyond traditional classifications, necessitating continuous adaptation of moderation policies to address emerging forms of harm. Similarly, Schöpke-Gonzalez and collaborators discussed how varying definitions of harm affect the annotation of data used in content moderation.[20] The study reveals that annotators do not use terms like "hateful," "offensive," and "toxic" interchangeably, indicating that perceptions of harmful content are subjective and evolving. This underscores the need for moderation policies to be flexible and responsive to changing societal attitudes. Furthermore, Scheuerman and collaborators presented a theoretical framework for understanding the severity of online harms. Their study identifies various types of harm and dimensions along which their severity can be assessed, highlighting the complex and evolving nature of harmful content. The authors suggest that content moderation policies must be adaptable to effectively address the shifting landscape of online harm.

40. As another example, research indicated that the COVID-19 pandemic led to a significant increase in anti-Asian hate speech on social media services. A study by researchers at Georgia Tech and Virginia Tech[21] analyzed over 200 million Tweets and found a notable rise in anti-Asian sentiment during the pandemic. The researchers observed that hate speech often surged following major COVID-19 news events, highlighting the dynamic nature of online hostility in response to global crises. Another study by researchers at CUNY examined the prevalence of anti-

[19] *See* Sheth, Amit, Valerie L. Shalin, and Ugur Kursuncu. *Defining and detecting toxicity on social media: context and knowledge are key*. Neurocomputing 490 (2022): 312-318.

[20] Schöpke-Gonzalez, Angela, et al. *How We Define Harm Impacts Data Annotations: Explaining How Annotators Distinguish Hateful, Offensive, and Toxic Comments*. arXiv preprint arXiv:2309.15827 (2023).

[21] He, Bing, et al. "Racism is a virus: Anti-Asian hate and counter-speech in social media during the COVID-19 crisis." Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining. 2021.

Asian hate speech on Twitter.[22] The findings revealed a significant increase in such content, particularly during the early stages of the pandemic.[23] Other studies further corroborated the escalation of anti-Asian sentiment on social media.[24] The pandemic intensified negative attitudes toward Asians, as evidenced by increased usage of racial slurs and derogatory language on services like Twitter. This trend is not limited to public health crises: a study by Olteanu and collaborators examined social media reactions following attacks involving Arabs and Muslims.[25] The research observed an increase in online hate speech, particularly messages advocating violence, directed at these communities after such events.

41. These studies underscore the need for effective detection and moderation strategies to address evolving online harms and collectively demonstrate that global events can significantly influence the emergence and prioritization of specific online harms, necessitating adaptive and responsive content moderation policies.

b. Ever-evolving regulatory challenges

42. The regulatory landscape for content moderation is complex and varies across jurisdictions, and this presents additional challenges for online services to perform content moderation. Of the thousands of regulations at the regional, county, and local level, some of the most significant regulatory regimes impacting content moderation efforts are described below:

---

[22] Toliyat, Amir, et al. "Asian hate speech detection on Twitter during COVID-19." Frontiers in Artificial Intelligence 5 (2022): 932381.

[23] To be sure, this content was by no means exclusive to social media platforms; in some circumstances, prominent politicians were fanning anti-Asian sentiments across other media platforms. *See, e.g.*, Itkowitz, C. (2020, Jan. 23). *Trump again uses racially insensitive term to describe coronavirus*. Washington Post. https://www.washingtonpost.com/politics/trump-again-uses-kung-flu-to-describe-coronavirus/2020/06/23/0ab5a8d8-b5a9-11ea-aca5-ebb63d27e1ff_story.html.

[24] Lu, Runjing, and Sophie Yanying Sheng. "How racial animus forms and spreads: Evidence from the coronavirus pandemic." Journal of Economic Behavior & Organization 200 (2022): 82-98.

[25] Olteanu, Alexandra, et al. "The effect of extremist violence on hateful speech online." Proceedings of the international AAAI conference on web and social media. Vol. 12. No. 1. 2018.

43. ***General Data Protection Regulation ("GDPR")***: The EU's GDPR imposes strict guidelines on the processing of personal data, emphasizing individual privacy rights. For content moderation, this means that services must carefully manage user data, particularly concerning data retention for AI model training.[26] The GDPR's principles of data minimization and purpose limitation can limit the extent to which user data is stored and utilized, potentially hindering further development and refinement of AI moderation tools.[27]

44. ***Digital Services Act ("DSA")***: The EU's 2022 DSA introduces comprehensive obligations for online services, including enhanced transparency requirements and stricter accountability for algorithmic content recommendations. Services are mandated to provide detailed explanations for content removal decisions and help ensure that their algorithms do not contribute to the dissemination of illegal content. This necessitates significant adjustments in content moderation practices to comply with the new standards.[28]

45. ***Children's Online Privacy Protection Act ("COPPA")***: The US's COPPA prohibits in certain cases the collection and retention of "personal information" from children under the age of thirteen without verifiable parental notice and consent.[29]

46. The scope of these and other regulatory regimes places significant hurdles before any social media service that moderates content, including Meta. Some of the most significant challenges impacting content moderation efforts are described below:

47. ***Limits to Human Review***: Privacy regulations may restrict the use of human reviewers in content moderation due to concerns over data protection. For instance, the GDPR

---

[26] *The impact of the GDPR on artificial intelligence*. (n.d.). Securiti.ai. Retrieved June 10, 2025, from https://securiti.ai/impact-of-the-gdpr-on-artificial-intelligence.

[27] *The impact of the GDPR on artificial intelligence*. (n.d.). Securiti.ai. Retrieved June 10, 2025, from https://securiti.ai/impact-of-the-gdpr-on-artificial-intelligence.

[28] Nunziato, D. C. (2023). The Digital Services Act and the Brussels Effect on Platform Content Moderation. Chi. J. Int'l L., 24, 115.

[29] Harmonizing laws to improve enforcement of human rights crimes. (n.d.). Thomson Reuters. Retrieved June 10, 2025, from https://www.thomsonreuters.com/en-us/posts/human-rights-crimes/harmonizing-laws/.

mandates that data processing involving human intervention must uphold user privacy rights, potentially limiting the extent to which human moderators can access and assess user content. This can challenge services' abilities to effectively manage nuanced content that AI systems may struggle to interpret.[30]

48. ***Restrictions on AI-Enhanced Moderation Tools***: Privacy laws like the GDPR may limit the ability to use AI-enhanced moderation tools that rely on extensive user data for training and improvement. The requirement for explicit user consent and the right to data deletion can reduce the data available for AI models, potentially diminishing their effectiveness in detecting harmful content.[31] In Europe, the Privacy Directive also prohibits processing of private messages without GDPR-level consent, also diminishing the data available for AI content moderation models.[32] Likewise, in the US, Meta has faced significant regulatory hurdles when developing AI technology that can reliably identify users under the age of thirteen at scale,[33] and COPPA's barriers to the collection and use of personal information from under-thirteen users have added to those challenges.[34]

49. ***Balancing Transparency and Security***: Transparency mandates, such as those in the DSA, require services to disclose information about their content moderation algorithms. While this promotes accountability, it also poses risks, as overly detailed disclosures could enable

---

[30] Information Commissioner's Office. (n.d.). What is content moderation and how does it use personal information?. Retrieved June 10, 2025, from https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/online-safety-and-data-protection/content-moderation-and-data-protection/what-is-content-moderation-and-how-does-it-use-personal-information/.

[31] *See* Masnick, M. (2018, January 30). Unintended consequences of EU's new internet privacy rules: Facebook won't use AI to catch suicidal users. Techdirt. https://www.techdirt.com/2018/01/30/unintended-consequences-eus-new-internet-privacy-rules-facebook-wont-use-ai-to-catch-suicidal-users/.

[32] *See* European Data Protection Board. (2020, May 4). *Guidelines 05/2020 on consent under Regulation 2016/679*. https://www.edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202005_consent_en.pdf; GDPR: Consent. (n.d.). Intersoft Consulting. Retrieved June 10, 2025, from https://gdpr-info.eu/issues/consent/.

[33] META3047MDL-065-00042982.

[34] META3047MDL-034-00027362.

malicious actors to game the system, circumventing moderation efforts. Services must carefully balance the need for transparency with the imperative to maintain the integrity of their moderation systems.

50. ***Tension Between Free Speech and Content Removal***: Regulations often require the removal of illegal or harmful content, but this can conflict with free speech protections, particularly in jurisdictions like the United States, where the First Amendment provides robust speech rights. Services must navigate these conflicting expectations, ensuring compliance with content removal mandates while respecting users' rights to free expression.[35]

51. ***Conflicting Government Expectations***: Different governments impose varying expectations on content moderation. The European Union, through regulations like the DSA, demands stricter moderation to prevent harm, whereas in the United States, there is often advocacy for less intervention to uphold free speech. This creates a complex regulatory environment for services operating globally, as they must tailor their moderation practices to comply with diverse legal frameworks.[15]

2. Technical Challenges for State of the Art Artificial Intelligence Systems

52. From a technical standpoint, the accuracy and effectiveness of content moderation systems in detecting and filtering out harmful content are ongoing subjects of research. Despite advances in natural language processing and image recognition technologies, algorithms still struggle with understanding context, sarcasm, and subtlety, leading to both over-censorship and underenforcement.[36]

---

[35] For example, in Thailand, laws exist that prohibit insulting the monarchy, which stands in stark contrast to US First Amendment protections associated with criticism of the government. BBC News. (2017, October 6). *Lese-majeste explained: How Thailand forbids insult of its royalty*. https://www.bbc.com/news/world-asia-29628191.

[36] *See* Gorwa, Robert, Reuben Binns, and Christian Katzenbach. *Algorithmic content moderation: Technical and political challenges in the automation of platform governance*. *Big Data & Society 7.1* (2020): 2053951719897945.

53. To be sure, artificial intelligence systems for text classification have advanced substantially in recent years, largely due to the emergence and adoption of transformer-based architectures. Introduced by Vaswani et al. in 2017,[37] the transformer model revolutionized natural language processing (NLP) by replacing traditional recurrent and convolutional networks architectures[38] with a self-attention mechanism that allows the model to weigh the relevance of each word in a sequence, regardless of its position. This design enables transformers to better capture long-range dependencies and contextual relationships within text.

54. Transformers form the backbone of modern large language models (LLMs), such as BERT (Bidirectional Encoder Representations from Transformers),[39] GPT (Generative Pre-trained Transformer),[40] and RoBERTa (a robustly optimized BERT variant),[41] which have set new benchmarks across a wide array of NLP tasks, including sentiment analysis, hate speech detection, and content moderation. These models are pre-trained on vast corpora of text using self-supervised objectives, and then fine-tuned on specific downstream tasks using labeled data. Their bidirectional (or autoregressive) architectures allow them to capture both left and right context, improving semantic understanding and disambiguation of meaning.

55. In the context of content moderation, transformer-based LLMs excel at parsing subtle linguistic signals such as sarcasm, idioms, code-switching, and culturally-specific

[37] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). *Attention is all you need*. Advances in neural information processing systems, 30.

[38] META3047MDL-208-00052347.

[39] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). *BERT: Pre-training of deep bidirectional transformers for language understanding*. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

[40] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*.

[41] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). *RoBERTA: A robustly optimized BERT pretraining approach*. arXiv preprint arXiv:1907.11692.

references that traditional machine learning models often miss.[42] Meta scientists adapted transformer-based models to their content moderation challenges, such as Meta's multilingual model like XLM-R,[43] to manage content across diverse languages and dialects. Furthermore, research by Meta's AI teams has led to the development of more efficient variants, such as the Linformer, which reduces the computational overhead associated with processing long sequences of text, making real-time moderation more scalable. [44]

56. Nevertheless, detecting and moderating harmful content on social media services like Meta's remains a multifaceted technical challenge. The main technical challenges associated with AI automated content moderation systems include the complexity of language, the scale and volume of content generated daily, resource intensiveness, identifying false positives and negatives, evolving adversarial threats (e.g., deepfakes), and the growing generation and alteration of existing content using AI.[45]

57. The technical challenges of detecting and moderating harmful content on social media services are continuously changing. Advances in AI, including those spearheaded by Meta's

[42] Chan, A. J., Redondo García, J. L., Silvestri, F., O'Donnell, C., & Palla, K. (2023). *Enhancing Content Moderation with Culturally-Aware Models*. arXiv e-prints, arXiv-2312.

[43] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

[44] Wang, S., Li, B. Z., Khabsa, M., Fang, H., & Ma, H. (2020). Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768.

[45] Gorwa, R., Binns, R., & Katzenbach, C. (2020). "Algorithmic content moderation: Technical and political challenges in the automation of platform governance." Big Data & Society.; Udupa, S., Maronikolakis, A., & Wisiorek, A. (2023). Ethical scaling for content moderation: Extreme speech and the (in) significance of artificial intelligence. Big Data & Society, 10(1), 20539517231172424; Huertas-García, Á., Martín, A., Huertas-Tato, J., & Camacho, D. (2023). Countering malicious content moderation evasion in online social networks: Simulation and detection of word camouflage. Applied Soft Computing, 145, 110552; Matias, J. Nathan. (2019). "The civic labor of volunteer moderators: Moderation, productivity, and the limits of automation at scale." New Media & Society, 21(6), 1271–1287.

AI scientists—such as efficient Transformer architectures and multimodal learning[46]—have significantly improved the ability to detect harmful content.

58. ***Cultural Challenges and Complexity of Language***: Cultural challenges arise when content moderation algorithms fail to recognize contextual nuances, leading to the suppression of culturally significant content or the inadvertent promotion of culturally insensitive material.[47] It is helpful for services to engage with experts from around the globe and incorporate a diverse set of perspectives to help ensure that algorithms are sensitive to cultural variations and do not enforce a one-size-fits-all approach to content moderation.[48] Automated systems must navigate the intricacies of human language, including sarcasm, slang, and cultural references, which vary across different languages and dialects.[49] This complexity makes it challenging for AI models to accurately interpret and classify content.[50] A recent study published by Matias and colleagues in *New Media & Society* highlights the severe limitations of existing machine learning content moderation methods in understanding such linguistic nuances: They found that classifiers often misinterpreted sarcasm, coded language, and in-group slang, flagging benign speech while failing to detect covert abuse. The study highlighted that models trained without cultural and contextual nuance performed poorly across diverse communities,[51] which can in turn lead to ethical concerns

[46] See *Wang, Sinong, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. "Linformer: Self-attention with linear complexity." arXiv preprint arXiv:2006.04768 (2020).* Note that this paper has been cited over a thousand times by researchers in the field of AI, substantiating its contribution and importance.

[47] *See* Gillespie, Tarleton. (2018.). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press

[48] META3047MDL-003-00001478.

[49] Indeed, the same holds true for human reviewers, as differences in cultural norms and experience may have a negative impact on the accuracy of human review.

[50] Udupa, S., Maronikolakis, A., & Wisiorek, A. (2023). Ethical scaling for content moderation: Extreme speech and the (in) significance of artificial intelligence. Big Data & Society, 10(1), 20539517231172424.

[51] Matias, J. Nathan. (2019). "The civic labor of volunteer moderators: Moderation, productivity, and the limits of automation at scale." New Media & Society, 21(6), 1271–1287.

about the fairness of such systems.[52]

59. Furthermore, there are subtleties of meaning in context, related to cultural or in-group communication norms that pose significant content moderation challenges: often, a word or image can assume different meanings in different communities. For instance, a picture of train tracks may have to do with travel (and therefore be innocuous content), or it may relate to SSI.[53] While humans who appreciate this context may more quickly tell the difference between these images, training AI to do so is challenging, particularly at scale. This is even harder on services like Instagram that are mostly image-based, where content often lacks any textual cues.[54]

60. For example, detecting hate speech and other harmful content requires a deep understanding of context, idioms, and cultural nuances, which even state-of-the-art models struggle to achieve consistently.[55] While attention weights offer some insight into model behavior, decisions made by these models are often opaque, complicating accountability and appeal processes in automated content enforcement.[56]

61. ***Multimodal and Multilingual Content Learning & Integration***: Harmful content is not confined to text alone – it can also appear in images, videos, audios, and a combination of these modalities. Additionally, it can be expressed in various languages[57] and dialects.[58] AI

[52] Udupa, S., Maronikolakis, A., & Wisiorek, A. (2023). Ethical scaling for content moderation: Extreme speech and the (in) significance of artificial intelligence. Big Data & Society, 10(1), 20539517231172424.

[53] META3047MDL-001-00000112.

[54] META3047MDL-001-00000112.

[55] Meta. (n.d.). *AI advances to better detect hate speech*. Retrieved June 10, 2025, from https://ai.meta.com/blog/ai-advances-to-better-detect-hate-speech/.

[56] Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018, April). 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In Proceedings of the 2018 Chi conference on human factors in computing systems (pp. 1-14).

[57] Haider, S., Luceri, L., Deb, A., Badawy, A., Peng, N., & Ferrara, E. (2023, April). Detecting social media manipulation in low-resource languages. In Companion Proceedings of the ACM Web Conference 2023 (pp. 1358-1364).

[58] Nicholas, G., & Bhatia, A. (2023). Toward better automated content moderation in low-resource languages. Journal of Online Trust and Safety, 2(1).

systems must integrate data from multiple modalities to accurately detect content that combines text and imagery, such as memes that use visual and textual elements to convey harmful messages.[59] Combining insights from different data modalities to form a coherent understanding of the content is a complex task that requires advanced algorithms and extensive training data.[60] Models must also be sensitive to the cultural contexts in which different languages are used to avoid misclassifying content due to cultural misunderstandings.[61] Costs to annotate multimodal data can rise rapidly at the scale that Meta operates.

62. ***Resource Intensiveness***: Transformer models are notoriously resource-intensive, requiring substantial computational resources and memory, which can limit their scalability. Costs to deploy and maintain these models can also rise rapidly at the scale Meta operates.[62] For example, Meta relies upon over 1,000 machine learning models at any one time, to continuously inspect content posted on Instagram.[63]

63. ***Scale and Volume***: Services like Facebook and Instagram process a colossal

[59] Sharma, S., Alam, F., Akhtar, M. S., Dimitrov, D., Martino, G. D. S., Firooz, H., ... & Chakraborty, T. (2022). Detecting and understanding harmful memes: A survey. arXiv preprint arXiv:2205.04274.

[60] Lippe, P., Holla, N., Chandra, S., Rajamanickam, S., Antoniou, G., Shutova, E., & Yannakoudakis, H. (2020). A multimodal framework for the detection of hateful memes. arXiv preprint arXiv:2012.12871.

[61] Bui, M. D., von der Wense, K., & Lauscher, A. (2024). Multi3Hate: Multimodal, Multilingual, and Multicultural Hate Speech Detection with Vision-Language Models. arXiv preprint arXiv:2411.03888.

[62] *See* Strubell, E., Ganesh, A., & McCallum, A. (2019). "Energy and policy considerations for deep learning in NLP." ACL. This study quantifies the high energy costs and carbon footprint of training transformer models, such as BERT and GPT-2, highlighting the practical trade-offs of deploying these models at large scale. State of the art models, such as OpenAI GPT4 and next-gen iterations are likely to have orders of magnitude higher costs. Fanatical Futurist. (2025, May). *OpenAI GPT-5 is costing $500 million per training run and still failing?* https://www.fanaticalfuturist.com/2025/05/openai-gpt-5-is-costing-500-million-per-training-run-and-still-failing/.

[63] Meta. (2025, May 21). *Journey to 1000 models: Scaling Instagram's recommendation system*. https://engineering.fb.com/2025/05/21/production-engineering/journey-to-1000-models-scaling-instagrams-recommendation-system/.

volume content daily,[64] requiring efficient algorithms and robust infrastructure for real-time data management: Meta's AI systems rely on hundreds of different statistical, algorithmic, and machine-learning models at any one time to process, filter, and recommend content.[65] The immense scale of data and the relentlessness of violations make AI approaches desirable, even inevitable, for content moderation. However, achieving effective moderation at this scale remains a significant challenge.[66] Despite algorithmic content moderation, Meta receives and processes millions of user reports each week.[67] The vast amount of content generated on social media services necessitates automated solutions for effective monitoring and management. AI systems must be highly accurate and efficient to operate in real-time across global services and varied data modalities (text, images, videos, audio). This requires continuous improvements in both hardware and software to maintain performance at scale.

64. ***Data Scarcity for Training***: Building effective AI models for content moderation requires extensive labeled data representing both harmful and benign content. Supervised learning (i.e., learning from examples, the dominant paradigm in AI moderation) relies on such annotations to train models that can generalize effectively to new inputs. However, certain types of violations, such as terrorism-related content, CSAM, or self-harm imagery, are relatively rare and legally restricted, making large-scale collection and labeling infeasible. Some scholars have highlighted how data imbalance and the scarcity of high-quality labeled examples limit the effectiveness of

[64] Smith, C. (2013, September 18). *Facebook users are uploading 350 million new photos each day*. Business Insider. https://www.businessinsider.com/facebook-350-million-photos-each-day-2013-9.

[65] Meta. (2025, May 21). *Journey to 1000 models: Scaling Instagram's recommendation system*. https://engineering.fb.com/2025/05/21/production-engineering/journey-to-1000-models-scaling-instagrams-recommendation-system/.

[66] Gorwa, Robert, Reuben Binns, and Christian Katzenbach. *Algorithmic content moderation: Technical and political challenges in the automation of platform governance*. *Big Data & Society 7.1* (2020): 2053951719897945.

[67] META3047MDL-003-00144612.

hate speech detection models.[68] Others similarly noted that content moderation models often fail to generalize across platforms or contexts due to lack of representative data, especially for low-resource languages and culturally specific slurs or euphemisms.[69] Platform policies, privacy law (e.g., GDPR), and ethical limitations introduce major constraints on the collection of training data involving vulnerable populations, such as children or marginalized users.[70] These constraints prevent the publication or open sharing of many datasets, limiting opportunities for cross-institutional model development or reproducibility. A useful case study is the Facebook Hateful Memes Challenge dataset (Kiela et al., 2020),[71] which combined image-text pairs to model subtle multimodal hate. Despite its innovation, the authors explicitly acknowledged limitations due to annotation complexity, cultural context dependence, and insufficient volume for broader generalization. This underscores how even large, well-designed datasets fall short of solving the scarcity challenge in real-world deployment. Another study on leveraging large-scale multimedia datasets highlighted the difficulty in creating accurate models due to the lack of adequate task-specific training data.[72] The authors demonstrated that even large, widely-used annotated datasets (such as the Twitter hate speech corpora frequently used to train content moderation systems,

[68] Schmidt, A., & Wiegand, M. (2017, April). A survey on hate speech detection using natural language processing. In Proceedings of the fifth international workshop on natural language processing for social media (pp. 1-10).

[69] Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. Plos one, 15(12), e0243300.

[70] Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 2053951719897945.

[71] Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., & Testuggine, D. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. Advances in neural information processing systems, 33, 2611-2624.

[72] Sarridis, I., Koutlis, C., Papadopoulou, O., & Papadopoulos, S. (2022, December). Leveraging large-scale multimedia datasets to refine content moderation models. In 2022 IEEE Eighth International Conference on Multimedia Big Data (BigMM) (pp. 125-132). IEEE.

including DWMW17,[73] and FDCL18,[74] etc.) suffer from systemic biases in how crowd workers label dialect-associated content, particularly African American English ("AAE"). These biases propagate into machine learning models trained on such data, leading to disproportionately high false positive rates for minority dialects. The study emphasizes that task-specific data is not only scarce but often misrepresentative, especially for marginalized users, and calls for more careful annotation protocols, dialect-awareness, and transparency in dataset design to ensure equity in content moderation.

65. ***Data Quality***: Ensuring the quality and representativeness of training data is essential to avoid biases and improve model performance. AI models can inherit biases present in training data, leading to unfair or discriminatory content moderation decisions. Biased algorithms can disproportionately affect certain demographics, reinforcing existing prejudices: for example, a prominent study empirically demonstrated how racial bias can be introduced and perpetuated in machine learning systems trained for content moderation.[75] The authors examined two widely-used Twitter datasets annotated for toxic language and uncovered a statistically significant correlations between features of AAE and labels for "offensive" or "abusive" content, finding that annotators (typically drawn from predominantly white crowd worker pools) were more likely to label Tweets written in AAE as toxic, even when those Tweets were benign or in-group vernacular. Left unaddressed, such biases can compromise fairness toward users. Biases can be mitigated by having humans working to ensure representativeness in training data, using dialect-primed or

[73] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In Proceedings of the international AAAI conference on web and social media (Vol. 11, No. 1, pp. 512-515).

[74] Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., ... & Kourtellis, N. (2018, June). Large scale crowdsourcing and characterization of twitter abusive behavior. In Proceedings of the international AAAI conference on web and social media (Vol. 12, No. 1).

[75] Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019, July). The risk of racial bias in hate speech detection. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 1668-1678).

demographically diverse annotation protocols, and auditing model outcomes across identity groups to prevent these harms. For example, Meta has made strides to collect demographic data with an aim toward improving the fairness of its technologies.[76]

66. ***Generalization***: Models must generalize from training data to real-world scenarios, which can vary significantly in terms of content and context. A recent study discusses the challenges in fine-tuning models to handle diverse and evolving content effectively.[77] The study explores these challenges using LLMs that are trained not just to label harmful content, but to explain why the content is harmful in plain language. This reasoning-based training helps the models better understand the meaning behind the words, which makes them more accurate and fair, even on content they have not seen before.

67. ***False Positives and False Negatives***: Balancing sensitivity and specificity in content moderation is particularly challenging because AI models must minimize false positives (i.e., incorrectly flagging benign content) and false negatives (i.e., failing to detect harmful content) among a sea of fluctuating contexts that make up human interaction. Continuous learning and model updates are necessary to improve detection accuracy. A report by *New America* emphasizes the complexity of automated tools in content moderation, noting the challenges that AI researchers face when they have to construct comprehensive datasets that account for the vast fluidity and variances in human language, leading to potential false positives and negatives.[78] For example, a recent study highlighted how two LLM-based models for content moderation based,

[76] Meta. (n.d.). *Assessing fairness of our products while protecting people's privacy*. Retrieved June 10, 2025, from https://ai.meta.com/blog/assessing-fairness-of-our-products-while-protecting-peoples-privacy/.

[77] Ma, H., Zhang, C., Fu, H., Zhao, P., & Wu, B. (2023). Adapting large language models for content moderation: Pitfalls in data engineering and supervised fine-tuning. arXiv preprint arXiv:2310.03400.

[78] New America. (n.d.). *The limitations of automated tools in content moderation*. Retrieved June 10, 2025, from https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/the-limitations-of-automated-tools-in-content-moderation.

ToxiGen-RoBERTa and RoBERTa-Toxicity-Classifier, based on work done by researchers at Google and Microsoft Research, which are widely adopted, exhibit different kinds of limitations that can affect false positive and negative results when classifying toxic content:[79]

- **Arbitrariness of model decisions**: The paper highlights that approximately 30% of English statements might be flagged as "toxic" (a potential false positive if it's legitimate content) by one version of a model, while another equally accurate version of the model might deem it acceptable.
- **Disparate impact**: The finding that arbitrary moderation decisions are unequally distributed across different demographic groups (e.g., anti-LGBTQ speech receiving more arbitrary decisions) implies that certain groups might be disproportionately affected by both false positives (legitimate speech being removed) and false negatives (harmful content targeting them being allowed).
- **Limitations in replicating human judgment**: The paper states that models can exhibit high levels of disagreement even on content where human annotators unanimously agree on its toxicity. This suggests that the models are either generating false negatives (missing clearly toxic content) or false positives (flagging content as toxic that humans agree is not), or both, at an inconsistent rate.

68. ***Evolving Digital Threats***: Malicious actors employ various strategies to evade detection. Strategies such as using deliberate misspellings or code words,[80] and altered images[81]

[79] Gomez, J. F., Machado, C., Paes, L. M., & Calmon, F. (2024, June). Algorithmic arbitrariness in content moderation. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 2234-2253).

[80] Wang, W., Huang, J. T., Wu, W., Zhang, J., Huang, Y., Li, S., ... & Lyu, M. R. (2023, May). Mttm: Metamorphic testing for textual content moderation software. In 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE) (pp. 2387-2399). IEEE.

[81] Jiang, Z., Zhang, J., & Gong, N. Z. (2023, November). Evading watermark-based detection of ai-generated content. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (pp. 1168-1181).

are commonly used to bypass AI filters. Spammers, propagandists, and other bad-faith users continually evolve their tactics to circumvent algorithmic detection, necessitating constant updates and improvements to integrity systems.[82] The adaptability of malicious actors poses a significant technical challenge.

69. For example, in recent years, Meta removed a network of malicious actors operating from Russia that relied on large networks of fake accounts that amplified each others' deceptive content.[83] Meta must constantly work to identify and stop coordinated efforts to bypass its content moderation systems.[84]

70. Adversarial actors often find ways to bypass integrity safeguards, employing tactics such as sophisticated bots, and coordinated inauthentic behaviors (e.g., state-sponsored influence operations, coordinated spam or fraud campaigns). A recent study analyzed how malicious users evade AI-based moderation on platforms like Facebook: [85] Drawing on a dataset of 500 flagged posts, the authors identified ten evasion tactics (including image alteration, language obfuscation, and contextual manipulation) that exploit weaknesses in automated moderation systems. The study concluded that while AI enables scalable moderation, it remains vulnerable to adversarial tactics and must be paired with human oversight, emphasizing the need for adaptive, context-aware

[82] See a) Ferrara, Emilio, et al. "The rise of social bots." Communications of the ACM 59.7 (2016): 96-104; b) Ferrara, Emilio. "The history of digital spam." Communications of the ACM 62.8 (2019): 82-91; c) Ferrara, Emilio. "GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models." Journal of Computational Social Science (2024): 1-21.l and d) Huertas-García, Á., Martín, A., Huertas-Tato, J., & Camacho, D. (2023). Countering malicious content moderation evasion in online social networks: Simulation and detection of word camouflage. Applied Soft Computing, 145, 110552.

[83] Meta. (2022, February). *January 2022 coordinated inauthentic behavior report*. Retrieved June 10, 2025, from https://about.fb.com/news/2022/02/january-2022-coordinated-inauthentic-behavior-report/.

[84] *E.g.*, Meta. (2022, February). *January 2022 coordinated inauthentic behavior report*. Retrieved June 10, 2025, from https://about.fb.com/news/2022/02/january-2022-coordinated-inauthentic-behavior-report/.

[85] Malec, L., & Lešetický, J. (2024). *Social media content moderation, censorship and AI detection evasion techniques*. IDIMT-2024: Changes to ICT, Management, and Business Processes through AI.

systems to respond effectively to evolving digital threats.

3. Limitations of human reviewers

71. ***Diverse Cultural and Linguistic Backgrounds***: One of the main challenges to robust and accurate human review on a global social media service is identifying and recruiting qualified human reviewers across the world who represent a spectrum of diverse cultures and languages. For example, Meta's services are used globally, requiring reviewers who understand local contexts, cultural nuances, and languages to accurately assess content.[86] What may be considered offensive or harmful in one culture might be acceptable in another. Ensuring that reviewers have this cultural sensitivity is crucial for fair and effective moderation.

72. ***Geographical Distribution***: A global social media service like Meta requires reviewers in different time zones to help ensure 24/7 coverage. This requires establishing review teams across various regions, which adds to the complexity of recruitment and management.

73. ***Scale***: Scaling up human review teams involves logistical complexities, such as recruiting, training, and managing a large and diverse workforce made up of around 35,000 content reviewers worldwide.[87] This requires significant investment in infrastructure and resources, and scaling up pools of reviewers for a particular market can sometimes take several months, if not longer (e.g., sub-Saharan African markets have proven difficult to scale).

74. ***Accuracy vs. Overenforcement***: Striking a balance between accurately detecting harmful content and avoiding the suppression of legitimate speech is crucial, as overly aggressive filters can stifle free expression and counterspeech.[88] Indeed, false positives of purportedly illegal content could have serious legal ramifications for wrongly accused users. In this regard, Meta's

---

[86] *See* Meta. (n.d.). *Community Standards*. Retrieved June 10, 2025, from https://transparency.meta.com/policies/community-standards/; Meta. (n.d.). *How review teams work*. Retrieved June 10, 2025, from https://transparency.meta.com/enforcement/detecting-violations/how-review-teams-work/.

[87] *See* META3047MDL-003-00144612.

[88] Meta. (n.d.). *Community Standards*. Retrieved June 10, 2025, from https://transparency.meta.com/policies/community-standards/.

posture is to strike a balance between restricting harmful content and promoting expression.[89]

* * *

75. Despite the swath of sociopolitical, technical, and logistical hurdles faced by social media services, Meta's content moderation policies and enforcement systems are industry-leading, comprehensive, and appropriately balanced as compared to the industry. To support this conclusion, I offer an explicit benchmarking framework, as in the following. First, I discuss Meta's iterative development of content moderation policies, which draw upon consultation with both internal stakeholders and external experts. Second, I explain how these policies are operationalized through Meta's moderation infrastructure, which combines machine learning classifiers, human review, and user reports. I also compare Meta's performance to the leading industry standards of the Digital Trust & Safety Partnership. In addition, I draw from my own academic and professional experience conducting audits and evaluations of large-scale moderation systems to further contextualize Meta's relative strengths.

IV. Meta's Content Moderation Policies Are Comprehensive and Appropriately Balanced

A. Overview and Development of Content Moderation Policy

76. Meta has long been at the forefront of advances in content moderation policy and systems to address and overcome, to the extent possible, the above-described challenges. As of July 2023, Meta has spent roughly $20 billion on user safety efforts since 2016, including on content moderation.[90] Before describing Meta's content moderation systems, I first provide an overview of Meta's comprehensive content moderation policies that direct the implementation of its systems: the public-facing Community Standards, the internal Policy Labs, and internal Borderline Content Policies, all of which govern Meta's content moderation systems.

77. Meta's policies are developed in conjunction with a panoply of internal and external

[89] META3047MDL-001-00000112.

[90] META3047MDL-113-00038326.

experts from around the globe who collaborate to develop objective policy rationales, to craft clear descriptions of Meta's policies for notice to the public, and to help continually refine policies and policy rationales over time as Meta enforces through its technologies.[91] Meta also collaborates on the operational side of content moderation, including by leveraging third-party signals and hashes, such that different proprietary content moderation systems communicate information about problematic content with each other and improve detection.[92] This feedback loop helps to ensure that Meta's policies continue to evolve with the realities of the technology on the ground.

78. Because of the amount of content on Meta's services, and because of Meta's careful and thoughtful approaches to content moderation, total removal of all policy-violating content is not possible. To get a sense of the scale, it is worth referring to Meta's latest available Full Year Disclosure 2024,[93] which reports that the Family daily active people (DAP)—measure that Meta utilizes to track how many active users are seen on Meta's family of apps on a daily basis—was 3.35 billion on average for December 2024. This was an increase of 5% year-over-year, suggesting that the service's user base continues to grow. The size of Meta's user base, at almost four billion active users, would make it the largest country in the world.[94] With this scale of a user base, there are several, sometimes competing priorities that inform Meta's policy decisions: chiefly safety, but also freedom of expression, privacy, and authenticity, among others.

79. Meta is uniquely situated to provide a service to foster free discussion,[95] and it

---

[91] *See discussion infra*.

[92] *E.g.*, Meta. (2019, August). *Open-source photo and video matching*. Retrieved June 10, 2025, from https://about.fb.com/news/2019/08/open-source-photo-video-matching/.

[93] Meta. (2025). *Meta reports fourth quarter and full year 2024 results*. Retrieved June 10, 2025, from https://investor.atmeta.com/investor-news/press-release-details/2025/Meta-Reports-Fourth-Quarter-and-Full-Year-2024-Results/default.aspx.

[94] Meta. (2023). *Meta reports fourth quarter and full year 2023 results; Initiates quarterly dividend*. Retrieved June 10, 2025, from https://s21.q4cdn.com/399680738/files/doc_financials/2023/q4/Meta-12-31-2023-Exhibit-99-1-FINAL.pdf.

[95] Scholars have referred to Facebook as the place where people gather to discuss issues that are vital to democracy. *See, e.g.* Heldt, A. P. M. (n.d.). Merging the social and the public: How

works to balance free discussion against its priority to provide "a safe environment for all who engage in acts of expression."[96] To encourage safe and free expression, Meta must follow a principled approach when removing content.

80. To this end, Meta has developed and employed a baseline set of public-facing rules known as the Community Standards that outline its general content moderation policies.[97] When Meta updates these public-facing policies, it publishes a change log showing the updates made in redline.[98]

81. The Community Standards do not go into extreme detail as to what constitutes a violation, and Meta does not otherwise disclose its internal guidelines that outline criteria for content removal, downranking (i.e., demoting such content in news feeds), and other content actions[99] because outlining its policy rationales and analytical frameworks for reviewing content could provide bad actors with a guideline to make content that skirts its policies.[100] Instead, more detailed internal guidelines, known as Policy Labs, lay out exactly what is and is not allowed on Meta's services.[101] Meta develops and maintains a highly detailed Policy Lab for each category of policy violating content relevant to this litigation. Each Policy Lab articulates a policy rationale, identifies known or common questions that arise during human review, and sets operational

---

social media platforms could be a new public forum. In Beyond the public square: Imagining digital democracy. Even the Supreme Court has referred to social media as the "modern public square." *Packingham v. North Carolina*, 137 S. Ct. 1730, 1737 (2017).

[96] Meta. (n.d.). *Community Standards*. Retrieved June 10, 2025, from https://transparency.meta.com/policies/community-standards/.

[97] Meta. (n.d.). *Community Standards*. Retrieved June 10, 2025, from https://transparency.meta.com/policies/community-standards/.

[98] *See, e.g.*, Meta. (n.d.). *Suicide and self-injury*. Retrieved June 10, 2025, from https://transparency.meta.com/policies/community-standards/suicide-self-injury/.

[99] *See infra* Part IV.

[100] *See, e.g.*, Meta. (2018, June). *Removing bad actors from Facebook*. Retrieved June 10, 2025, from https://about.fb.com/news/2018/06/removing-bad-actors-from-facebook/.

[101] META3047MDL-177-00000003; META3047MDL-177-00000047; META3047MDL-177-00000119; META3047MDL-177-00000170; META3047MDL-203-00203440 .

guidelines.[102] Each Policy Lab also provides detailed examples of how certain content was actioned. Further, Meta has an internal set of guidelines governing the review and enforcement against "borderline content."[103] Borderline content is that which "does not explicitly violate" Meta's Community Standards but which "may come close."[104] Meta recognizes that universal content policies come with inherent challenges, including that "some people have a definition of what feels problematic that goes beyond what" Meta removes through its Community Standards."[105] Getting the balance right requires "a range of expertise," including policy, product, and operations teams.[106]

82. Internally, many Meta teams are focused on developing and refining content moderation policy. Meta's Content Policy, Integrity, Global Operations, and Research teams collaborate to maintain Meta's Policies to reflect the most up-to-date information with respect to the parameters of harmful content and the relative risk it poses at any given time.[107] The Content Policy team, which sits in more than a dozen locations around the world, is responsible for developing Meta's Community Standards and Community Guidelines.[108] The team includes subject-matter experts across many harm types.[109] The Integrity team then assesses the global impact of the potential policy change and builds the technology to scale the detection and

[102] META3047MDL-177-00000003; META3047MDL-177-00000047; META3047MDL-177-00000119; META3047MDL-177-00000170; META3047MDL-203-00203440.

[103] META3047MDL-113-00000439.

[104] META3047MDL-113-00000439.

[105] META3047MDL-113-00000439.

[106] META3047MDL-113-00000439.

[107] Meta. (n.d.). *Deciding to change standards*. Retrieved June 10, 2025, from https://transparency.meta.com/policies/improving/deciding-to-change-standards/.

[108] Meta. (n.d.). *Deciding to change standards*. Retrieved June 10, 2025, from https://transparency.meta.com/policies/improving/deciding-to-change-standards/.

[109] Meta. (n.d.). *Deciding to change standards*. Retrieved June 10, 2025, from https://transparency.meta.com/policies/improving/deciding-to-change-standards/.

enforcement of the updated policies.[110] The Global Operations team (f/k/a/ the Community Operations team), which is responsible for partnering with outside partners and contractors to enforce Meta's Policies, keeps Meta informed of trends in enforcement so that Meta can either clarify its policies and/or upgrade its technology.[111] Further, Meta consults regularly with experts in adolescent development, psychology, and mental health to help ensure its services are safe and age-appropriate.[112]

83. At biweekly "Product Policy Forums," Meta convenes a variety of subject-matter experts from its "safety and cybersecurity policy teams, counterterrorism specialists, Global Operations employees, product managers, public policy leads and representatives from our legal, communications and diversity teams."[113] Meta convenes such forums to discuss potential policy issues and propose policy changes, including but not limited to content moderation policies. In the interest of transparency, the minutes from each forum are published online.[114] Policy proposals include perspectives gathered from discussions with academics, NGOs, or other experts and stakeholders outside the company. In the case of policies addressing SSI content, Meta has gathered perspectives from the National Eating Disorders Association, National Suicide Prevention Lifeline, Crisis Text Line, Samaritans, beyondblue, headspace, the Duke Center for Eating Disorders, victims, and others.[115]

---

[110] Meta. (n.d.). *Deciding to change standards*. Retrieved June 10, 2025, from https://transparency.meta.com/policies/improving/deciding-to-change-standards/.

[111] Meta. (n.d.). *Deciding to change standards*. Retrieved June 10, 2025, from https://transparency.meta.com/policies/improving/deciding-to-change-standards/.

[112] Meta. (2024, January). *New protections to give teens more age-appropriate experiences on our apps*. Retrieved June 10, 2025, from https://about.fb.com/news/2024/01/teen-protections-age-appropriate-experiences-on-our-apps/.

[113] Meta. (2018, November). *Product policy forum minutes*. Retrieved June 10, 2025, from https://about.fb.com/news/2018/11/content-standards-forum-minutes/.

[114] Meta. (2018, November). *Product policy forum minutes*. Retrieved June 10, 2025, from https://about.fb.com/news/2018/11/content-standards-forum-minutes/.

[115] META3047MDL-003-00081293; META3047MDL-003-00191358.

B. Summary of Public-Facing Content Moderation Policies on Facebook or Instagram

1. Suicide & Self-Injury & Eating Disorder Content

84. First, a summary of the key aspects of Meta's public-facing policies on suicide, self-injury, and eating disorder content is provided in the following policy card.[116]

Table 1. Meta Policy Card on SSI & ED

Policy Card	Details
Policy Name	Suicide, Self-Injury, and Eating Disorders
Policy Rationale	We care deeply about the safety of the people who use our apps. We regularly consult with experts in suicide, self-injury and eating disorders to help inform our policies and enforcement, and we work with organizations around the world to help people in distress. While we do not allow people to intentionally or unintentionally celebrate or promote suicide, self-injury or eating disorders, we do allow people to discuss these topics because we want our services to be a space where people can share their experiences, raise awareness about these issues, and seek support from one another.
Allowed Content	- Discussion of suicide, self-injury, or eating disorders aimed at sharing experiences, raising awareness, or seeking support. - Content about recovery that may contain non-graphic imagery (e.g., healed scars) placed behind a sensitivity screen.
Restricted Content	- Content that promotes, encourages, coordinates, or provides instructions for suicide, self-injury, or eating disorders. - Graphic self-injury imagery. - Depictions of a person who engaged in a suicide attempt or death by suicide. - Mocking victims or survivors of suicide, self-injury, or eating disorders. - Imagery depicting body modification in a suicide or self-injury context.
Content Labels and Restrictions	- Warning screens for sensitive content (e.g., healed cuts in a recovery context). - Age restriction (18 and older) for content depicting euthanasia/assisted suicide.
Enforcement	- Regular consultation with external experts. - Contacting emergency services for immediate risks.

85. In addition to its public-facing policies, Meta maintains an internal SSI & ED

[116] Meta. (n.d.). *Suicide, self-injury, and eating disorders.* Meta Transparency Center. Retrieved May 4, 2025, from https://transparency.meta.com/policies/community-standards/suicide-self-injury/.

Policy Lab.[117] While Meta has one Policy Lab that covers both SSI and ED content, these harm types are addressed separately. In the SSI portions of this Policy Lab, Meta distinguishes between different types of imagery depicting graphic suicide content. For example, Meta distinguishes between imagery that depicts a suicide attempt or death by suicide, on the one hand, and imagery that depicts a person engaging in euthanasia or physician-assisted suicide, on the other. While the former would be removed, the latter would be allowed but placed behind a sensitivity screen including a warning that the content may be disturbing. In the self-injury context, the Policy Lab distinguishes between content that depicts graphic self-injury and content that depicts healed cuts in the context of recovery.

86. To help ensure that its SSI & ED Policy Lab is robust and effective, Meta has continuously consulted with mental health experts and organizations such as Forefront, Now Matters Now, Save.org, and the National Suicide Prevention Lifeline.[118] These consultations are key because SSI content creates a unique conundrum in the context of content moderation: auto-enforcement of all SSI content "could overenforce on people that are making cries for help" as "[s]afety professionals mostly believe that SSI admission should provide resources to the users and not punitive actions.[119] Consistently, experts have told Meta that sometimes, certain SSI content can be shared in a positive context and can help to destigmatize mental health struggles.[120] As explained by Dr. Reidenberg (Psy.D, FAPA, Executive Director, Save.org), whom Meta has consulted in the development of its policy approach to SSI content:

---

[117] META3047MDL-177-00000003.

[118] META3047MDL-001-00000112.

[119] META3047MDL-003-00079819.

[120] META3047MDL-001-00000112 . One stakeholder explains, "Mental illness and thoughts of suicide are just not something we talk about OPENLY. Yet talking and connecting is crucial to helping prevent depression and suicide. The tools Facebook is rolling out, aim both at people who are expressing suicidal thoughts and also guide concerned friends or family members to resources and alternatives and appropriate interventions." Meta. (n.d.). *Suicide prevention*. Retrieved June 10, 2025, from https://about.meta.com/actions/safety/topics/wellbeing/suicideprevention.

> When someone expresses suicidal distress, it provides family, friends and even Facebook and Instagram the opportunity to intervene. If people can't share their pain, or it is shared and then removed, we've missed a chance to save someone's life. We train people to listen for this in conversations and to allow people to keep talking because we know that it is one way to help them through a crisis. Social media services allow us to do this in a way that brings many people to help very quickly.[121]

87. Furthermore, Meta recently collaborated with Snapchat on an initiative called Thrive, through which the companies have shared SSI-related signals. Meta provided the technical infrastructure for this initiative.[122]

88. The SSI & ED Policy Lab establishes highly nuanced implementation standards for ED content. These standards are designed to help ensure that violating content is actioned while at the same time preserving the important role that Meta services play in promoting recovery and connecting users with critical support. For example, Meta removes imagery depicting ribs, collarbones, thigh gaps, hips, concave stomach, or protruding spine, unless that imagery is shared in the context of recovery from an eating disorder.[123] But, when such content is shared in the context of recovery, Meta places it behind a sensitivity screen rather than removing the content.[124]

89. The SSI & ED Policy Lab also contains a glossary of terms that are relevant to eating disorder content. For example, the glossary defines the term "long-shot" as "imagery where part of the depicted person's shoulder or knee appears in the frame."[125] The Policy Lab also addresses questions that Meta knows arise in the context of human review of eating disorder content, such as which principles or criteria are considered when deciding whether a term or concept is promoting or signaling eating disorders. The answers list hashtags that are associated

---

[121] Meta. (n.d.). *Suicide prevention*. Retrieved June 10, 2025, from https://about.meta.com/actions/safety/topics/wellbeing/suicideprevention.

[122] Meta. (2024, September). *Preventing suicide and self-harm content from spreading online*. https://about.fb.com/news/2024/09/preventing-suicide-and-self-harm-content-spreading-online/.

[123] META3047MDL-177-00000003.

[124] META3047MDL-177-00000003.

[125] META3047MDL-177-00000003.

with eating disorders, such as "thinspo, bonespo, proana, promia, and thighgap."[126]

2. Bullying and Harassment Content

90. Next, a summary of the key aspects of Meta's public-facing policies on bullying and harassment ("BH") is provided in the following policy card:

Table 2. Meta Policy Card on Bullying & Harassment[127]

Policy Card	Details
Policy Name	Bullying and Harassment
Policy Rationale	Bullying and harassment happen in many places and come in many different forms: from making threats and releasing personally identifiable information to sending threatening messages and making unwanted malicious contact. We do not tolerate this kind of behavior because it prevents people from feeling safe and respected on Facebook, Instagram, and Threads. We distinguish between public figures and private individuals because we want to allow discussion, which often includes critical commentary of people who are featured in the news or who have a large public audience. For public figures, we remove attacks that are severe as well as certain attacks where the public figure is directly tagged in the post or comment. For private individuals, our protection goes further: We remove content that's meant to degrade or shame, including, for example, claims about someone's sexual activity. We recognize that bullying and harassment can have more of an emotional impact on minors, which is why our policies provide heightened protection for anyone under the age 18, regardless of user status.
Allowed Content	- Content that condemns or draws attention to bullying and harassment. - Discussion of public figures with critical commentary, if it does not include severe attacks or direct tagging.
Restricted Content	- Unwanted contact that is repeated, sexually harassing, or directed at many individuals with no prior solicitation. - Calls for self-injury or suicide. - Attacks based on experiences of sexual assault, exploitation, harassment, or domestic abuse. - Severe sexualized commentary and derogatory terms related to sexual activity. - Threats to release private information. - Content that degrades or expresses disgust towards individuals depicted in processes like menstruating, urinating, vomiting, or defecating.

[126] META3047MDL-177-00000003.

[127] Meta. (n.d.). *Bullying and harassment*. Meta Transparency Center. Retrieved May 4, 2025, from https://transparency.meta.com/policies/community-standards/bullying-harassment/.

Policy Card	Details
Content Labels and Restrictions Enforcement	- Enhanced protections for minors under age 18. - Warning screens for sensitive content. - Regular consultation with external experts. - Bullying Prevention Hub for resources and support.

91. The BH Policy Lab establishes a series of "universal protections" that apply to all users, as well as more specific protections that depend on the type of user involved.[128] For example, Meta applies stricter protections when the user is a minor.[129] Similarly, Meta's protections go further when the user is a private individual, as opposed to a public figure.[130]

92. The BH Policy Lab is designed to address the context and intent behind a piece of content before actioning it. This helps Meta target bullying and harassment without inadvertently deleting content that was shared to condemn or draw attention to bullying and harassment, for example.

93. The first tier of protections, the universal protections, apply across all user types and across all types of content (e.g., organic content, paid/ads).[131] These universal protections target, among other things, unwanted contact that is repeated, sexually harassing, or that calls for self-injury or suicide of a specific person or group. These universal protections also apply to content that threatens to release an individual's private information.[132]

94. The second tier of protections apply to all minors, private adults, and any limited-scope public figures.[133] Meta removes content that includes claims about sexual activity, except for claims that are in the context of criminal allegations against an adult. Meta also removes any content that highlights or otherwise draws negative attention to a person's specific characteristics.

[128] META3047MDL-177-00000047.
[129] META3047MDL-177-00000047.
[130] META3047MDL-177-00000047.
[131] META3047MDL-177-00000047.
[132] META3047MDL-177-00000047.
[133] META3047MDL-177-00000047.

95. The third tier sets out additional protections for private minors, private adults, and others.[134] Where this tier applies, Meta removes content such as claims about someone's sexual orientation or content that expresses contempt or disgust about an individual. Many of these terms, such as contempt or disgust, are defined in the Policy.

96. The fourth tier sets out additional protections for private minors, such as removing any videos of physical bullying against a minor shared in a condemning context.[135]

97. In addition to the tiers described above, the BH Policy Lab sets out specific exceptions and allowances.[136] For example, though Meta would ordinarily remove videos of physical bullying, it allows such videos in the context of fighting sports such as martial arts. Furthermore, the Policy Lab sets out standards according to which Meta can identify the individual being targeted in a bullying and harassing post. For example, identifying relevant responses from the potential target in the comments when the parent post does not explicitly call out the target by name or image; or the use of pronouns or the mentioning of an individual without use of their name, along with comments indicating that the target referred to by the pronoun is known to the people engaging with the post.[137] This is a complex task since bullies often do not name their target but rather describe them (*e.g.*, "this girl I work with . . . .").

98. Finally, though the Policy Lab addresses a wide range of content, Meta also allows users to submit reports to the service. This is particularly useful in the BH context because the target may subjectively feel that they are being bullied or harassed, even if the content does not fit into any of the categories outlined in the Policy.

3. Adult Nudity & Sexual Activity Content

99. A summary of the key aspects of Meta's public-facing policies on adult nudity and sexual activity content is provided in the following policy card.

---

[134] META3047MDL-177-00000047.

[135] META3047MDL-177-00000047.

[136] META3047MDL-177-00000047.

[137] META3047MDL-177-00000047.

Table 3. Meta Policy Card on ANSA[138]

Policy Card	Details
Policy Name	Adult Nudity and Sexual Activity
Policy Rationale	The display of nudity or sexual activity is restricted due to sensitivities related to cultural backgrounds and age. Exceptions are made for content intended for protest, awareness, educational, or medical reasons.
Allowed Content	- Images of female breasts in protest, breastfeeding, and post-mastectomy scarring. - Artistic depictions, such as paintings and sculptures, that include nude figures. - Medical or health-related imagery involving visible genitalia, anuses, or female nipples.
Restricted Content	- Photorealistic imagery of adult nudity, including visible genitalia, anuses, and female nipples, except in specific contexts. - Explicit sexual activity, including intercourse, oral sex, and the use of sex toys. - Implicit sexual activity unless in medical, health, or recognized fictional contexts. - Fetish-related activities that could lead to harm, including bestiality and incest. - Digital imagery of adult sexual activity unless it is for medical awareness, scientific discourse, or sexual health discussions.
Content Labels and Restrictions Enforcement	- Sensitive content labels are added to certain images to inform viewers. - Viewing of some content is restricted to users aged 18 and older. - Global review teams and external stakeholders help enforce and inform the policy. - Data on prevalence, content actioned, proactive rate, appealed content, and restored content is tracked and reported.

100. The ANSA Policy Lab also establishes universal protections that nevertheless give way to exceptions for nuanced situations.[139]

101. ANSA content is prohibited outright in "ads and commerce surfaces."[140] This includes "nudity, depictions of people in explicit or suggestive positions, or activities that are

[138] Meta. (n.d.). Adult nudity & sexual activity. Meta Transparency Center. Retrieved May 4, 2025, from https://transparency.meta.com/policies/community-standards/adult-nudity-sexual-activity/.

[139] META3047MDL-177-00000170.

[140] META3047MDL-177-00000170.

overly sexually suggestive or provocative."[141]

102. For organic (i.e., non-ad or non-commercial content), Meta "default[s] to removing sexual imagery to prevent the sharing of non-consensual or underage content."[142]

103. In other contexts, "where appropriate and [] intent is clear," Meta may allow ANSA content shared "as a form of protest, to raise awareness about a cause or for educational or medical reasons."[143] Nevertheless, most ANSA content that is not defaulted to removal is filtered for users under the age of eighteen or marked as sensitive.[144]

4. Violent & Graphic Content

104. A summary of the key aspects of Meta's public-facing policies on violent and graphic ("VG") content is provided in the following policy card.

Table 4. Meta Policy Card on VG[145]

Policy Card	Details
Policy Name	Violent and Graphic Content
Policy Rationale	To protect users from disturbing imagery, we remove content that is particularly violent or graphic, such as videos depicting dismemberment, visible innards or charred bodies. We also remove content that contains sadistic remarks towards imagery depicting the suffering of humans and animals. In the context of discussions about important issues such as human rights abuses, armed conflicts or acts of terrorism, we allow graphic content (with some limitations) to help people to condemn and raise awareness about these situations.
Allowed Content	- Graphic content related to human rights abuses, armed conflicts, or acts of terrorism with some limitations. - Content with warning labels for graphic or violent imagery.

[141] META3047MDL-177-00000170.

[142] META3047MDL-177-00000170.

[143] META3047MDL-177-00000170.

[144] META3047MDL-177-00000170.

[145] Meta. (n.d.). Violent & graphic content. Meta Transparency Center. Retrieved May 4, 2025, from https://transparency.meta.com/policies/community-standards/violent-graphic-content/.

Policy Card	Details
Restricted Content	- Videos of dismemberment, visible internal organs, charred bodies, and victims of cannibalism. - Sadistic remarks towards disturbing imagery. - Live streams of capital punishment. - Graphic imagery of animal suffering.
Content Labels and Restrictions	- Warning labels for graphic or violent imagery. - Age restriction (18 and older) for certain graphic content. - Regular consultation with external experts.
Enforcement	- Specific measures for removing violent death videos upon request by family members.

105. Meta's VG Policy Lab protects users from sensitive content while also accounting for situations in which such content may serve a purpose, such as advocating against human rights abuses.[146] This Policy Lab is designed with the understanding that different people have different sensitivities to violent and graphic imagery. As such, Meta frequently adds warning screens to content instead of removing it entirely.

106. Furthermore, the VG Policy Lab accounts for situations in which graphic content may be acceptable.[147] For example, while content showing visible innards is generally not permitted, Meta does allow visible innards in the birthing context, though such content would be marked as sensitive.[148] Similarly, even if a video shows violent death or a life-threatening event, Meta may stop short of removing it because of the context of the video.[149] For example, Meta will mark as sensitive a video showing violence committed by uniformed police. And, while content depicting punctured skin may be removed in some contexts, it is generally allowed in the context of vaccines or piercings.[150]

107. Meta has a slightly different set of policies for fictional graphic imagery, and these

[146] META3047MDL-177-00000119.
[147] META3047MDL-177-00000119.
[148] META3047MDL-177-00000119.
[149] META3047MDL-177-00000119.
[150] META3047MDL-177-00000119.

policies distinguish between photorealistic and non-photorealistic imagery.[151] Generally, Meta is more permissive of violent or graphic content when it is non-photorealistic.

5. Child Safety Content

108. A summary of the key aspects of Meta's public-facing policies on child safety content is provided in the following policy card.

Table 5. Meta Policy Card on

Child Sexual Exploitation, Abuse, and Nudity ("Child Safety" or "CS") [152]

Policy Name	Child Safety Content
Policy Rationale	We do not allow content or activity that sexually exploits or endangers children. When we become aware of apparent child exploitation, we report it to the National Center for Missing and Exploited Children (NCMEC), in compliance with applicable law. We know that sometimes people share nude images of their own children with good intentions; however, we generally remove these images because of the potential for abuse by others and to help avoid the possibility of other people reusing or misappropriating the images.
Allowed Content	Content from real-world art, cartoons, movies, or video games that depicts real or non-real non-sexual child abuse.
Restricted Content	- Content, activity, or interactions that threaten, depict, praise, support, provide instructions for, make statements of intent, admit participation in, or share links of the sexual exploitation of children (including real minors, toddlers, or babies, or non-real depictions with a human likeness, such as in art, AI-generated content, fictional characters, dolls, etc.). - Content that solicits sexual content or activity depicting or involving children - Content that constitutes or facilitates inappropriate interactions with children - Content that attempts to exploit real children - Content (including photos, videos, real-world art, digital content, and verbal depictions) that sexualizes real or non-real children - Groups, Pages, and profiles dedicated to sexualizing real or non-real children

[151] META3047MDL-177-00000119.

[152] Child Sexual Exploitation, Abuse, and Nudity (n.d.). Meta Transparency Center. Retrieved on May 4, 2025 at https://transparency.meta.com/policies/community-standards/child-sexual-exploitation-abuse-nudity/.

Policy Name	Child Safety Content
	- Content that depicts real or non-real child nudity - Videos or photos that depict real or non-real non-sexual child abuse regardless of sharing intent, unless the imagery is from real-world art, cartoons, movies or video games - Content that praises, supports, promotes, advocates for, provides instructions for or encourages participation in non-sexual child abuse
Content Labels and Restrictions and Enforcement	For the following content, we include a warning screen so that people are aware the content may be disturbing and limit the ability to view the content to adults ages eighteen and older: - Videos or photos that depict police officers or military personnel committing non-sexual child abuse - Videos or photos of non-sexual child abuse, when law enforcement, child protection agencies, or trusted safety partners request that we leave the content on the service for the express purpose of bringing a child back to safety. For the following content, we include a sensitivity screen so that people are aware the content may be upsetting to some: - Videos or photos of violent immersion of a child in water in the context of religious rituals For the following content, we include a warning label so that people are aware that the content may be sensitive: - Imagery posted by a news agency that depicts child nudity in the context of famine, genocide, war crimes, or crimes against humanity, unless accompanied by a violating caption or shared in a violating context, in which case the content is removed

109. Reinforcing Meta's Community Standards, Meta's CS Policy Lab establishes universal protections that leave little room for exceptions and direct the removal of the overwhelming majority of CS content, with restrictions and limitations applied to the rest. The Policy Lab's Operational Guidelines, which human reviewers reference when reviewing potentially policy-violating content, contain highly granular factors for reviewers to consider.[153] For example, when determining minor age, reviewers are directed to consider explicit but also indirect information, such as the relative size of body parts, the relative roundness of faces, and the relative lack of a chin or jawline.[154]

[153] META3047MDL-203-00203440.
[154] META3047MDL-203-00203440.

C. Borderline Policies

110. Further to its policies regarding clearly violating content, I understand Meta has had internal "borderline content" policies since at least 2020.[155] Borderline Content Policies address content that exists in a grey area, i.e., content that does not explicitly violate the Community Standards but may still be considered harmful or controversial.[156] These policies are designed to manage and mitigate the spread of such content, helping to ensure that the service remains a safe and welcoming space for all users. Borderline content includes material that skirts the edge of what is acceptable under Meta's Community Standards.[157] This might include content that is sensationalist, misleading, or provocative but not outright false or harmful.

111. For example, "Meta has an SSI Borderline Policy that covers things that are close to prohibited, or SSI-adjacent like dark and depressing content."[158] Meta's ED Borderline Policy is designed "to limit user exposure to content posted by non-connected users that could cause harm when viewed either by a vulnerable user or as part of an aggregate viewing experience."[159] Meta's VG Borderline Policy covers "things like pus or oozing wounds, people undergoing surgical operations, fight videos, animal abuse, and fictional violence."[160] Finally, Meta's ANSA Borderline Policy covers "things like implied sexual intercourse and stimulation."[161] Meta does not have Borderline Policies for CS content because the Community Standards and CS Policy Lab

[155] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 30, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[156] META3047MDL-113-00000439.

[157] META3047MDL-113-00000439.

[158] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 31, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[159] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 31, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[160] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 31, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[161] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 31, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

are the strictest set of content moderation policies and permit almost no, if any, nuance in the moderation of CS content.[162]

112. Because borderline content is not policy violating, Meta does not remove borderline content but does apply various "soft" enforcement actions.[163] "Soft action" refers to an action that Meta takes on a piece of content short of removal to reduce its exposure to Meta's users, such as downranking or demoting, filtering, age gating, adding warning screens or captions, non-recommendations, in-feed recommendation (IFR) filtering, or search engine results page (SERP) filtering.[164] Training machine learning models to detect borderline content is technically challenging and complex because it deals with content less amenable to categorization and assessment of the trade-offs between safety and voice.[165]

V. Meta's Content Moderation Systems Are Robust and Reasonably Designed.

113. Meta utilizes a number of different systems to monitor and detect prohibited content and profiles on Facebook and Instagram. Enforcement is governed by three prongs:[166]:

114. ***Removal***: Meta removes content and, in certain instances, accounts that violate its policies. Users have an opportunity to appeal the removal.

115. ***Reduction***: If content does not violate Meta's content moderation policies but is still deemed sensitive, or borderline, Meta reduces its distribution and/or visibility generally or for a specific subset of users (e.g., users under the age of eighteen). Certain categories of content are barred (i.e., age gated) altogether from distribution or visibility to users under the age of eighteen. Meta also provides users with additional tools to opt in to certain automated content moderation features, such as hiding certain words.

---

[162] See Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[163] *See infra* Appendix A.

[164] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 7, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[165] *See* META3047MDL-113-00000439.

[166] META3047MDL-001-00000112 .

116. ***Information***: Meta seeks to provide context on sensitive content before a user is exposed to it (such as via warning screens).

117. This section proceeds in two parts. In the first, I will explain how Meta's robust proactive detection and review mechanisms work and detail the specific mechanisms in place for different types of harmful content (e.g., SSI, ED, BH, VG, ANSA, and CS). In the second part, I will explain how Meta's network of human reviewers supports its proactive detection and review mechanisms.

A. Proactive Detection

118. Meta identifies and actions most policy-violating content proactively, rather than depending on users to report it. Meta uses AI tools to review and categorize content that may need to be removed, downranked, or otherwise actioned.[167] One of the most important sets of tools are "classifiers." Classifiers assess, detect, and categorize content, including text, video, and photos, on both Facebook and Instagram to determine whether the content violates Meta's Community Standards.[168] In addition, Meta has a suite of other automated features that help reduce user exposure to prohibited content or abusive accounts.

1. Meta's Early Leadership in Proactive Detection Technology

a. Facebook Immune System

119. Meta has pioneered and iterated upon content moderation systems for many years, beginning with the Facebook Immune System ("FIS"), launched by early 2011.[169] FIS was a first solution for a real-time, large-scale integrity system and served as an early exemplar of adversarial machine learning systems tailored for dynamic, high-volume environments. FIS was a cornerstone

[167] Meta. (n.d.). *How our enforcement technology works*. Retrieved June 10, 2025, from https://transparency.fb.com/enforcement/detecting-violations/how-enforcement-technology-works/.

[168] *See supra* Part IV; *see also infra* Appendix A (listing the topline classifiers used on Facebook and Instagram as of February 2025).

[169] Stein, T., Chen, E., & Mangla, K. (2011, April). *Facebook immune system*. In Proceedings of the 4th workshop on social network systems. https://css.csail.mit.edu/6.858/2014/readings/facebook-immune.pdf.

in Meta's defense infrastructure against adversarial threats such as phishing, spam, and abuse across its social services.[170]

120. At its core, FIS was a system designed to detect and mitigate threats such as spam, fake accounts, misinformation, and other forms of abuse that can harm the integrity of the platform. It did this by continuously monitoring and classifying billions of user interactions daily, including reads and writes, to flag potential issues. It took automated actions such as removing content, disabling accounts, or alerting human reviewers for further investigation to proactively defend the social graph from abuse. FIS, initially designed to combat security threats, later evolved into a key content moderation engine, enabling scalable, automated enforcement of policies related to harmful content, misinformation, hate speech, and platform manipulation. FIS utilized a variety of machine learning models and algorithms that analyze user behavior, content, and interactions to identify suspicious activities. These models were continuously updated to adapt to new threats. Further, FIS enabled real-time moderation of user-generated content with latencies under 50 milliseconds.[171]

121. Moderation decisions were informed by both explicit signals (e.g., "mark as spam") and implicit behaviors (e.g., post deletions, rejected friend requests), which serve as training data and anomaly triggers. Classifier services ran various ML algorithms (e.g., random forests, SVMs) to evaluate content risk in real-time. Feature Extraction Language (FXL) enables on-the-fly creation and deployment of new moderation features without system downtime. Policy Engine translates classifier scores into enforcement actions (e.g., content takedown, account restriction) via flexible, rule-based logic. Further, FIS aggregated features across multiple communication channels (posts, messages, comments, friend requests) to detect violations that span modalities or

[170] Stein, T., Chen, E., & Mangla, K. (2011, April). *Facebook immune system*. In Proceedings of the 4th workshop on social network systems. https://css.csail.mit.edu/6.858/2014/readings/facebook-immune.pdf.

[171] Stein, T., Chen, E., & Mangla, K. (2011, April). Facebook immune system (pp. 2–3). In Proceedings of the 4th workshop on social network systems. https://css.csail.mit.edu/6.858/2014/readings/facebook-immune.pdf

contexts.

122. In summary, FIS represented one of the earliest large-scale, adversarially aware content moderation systems and laid the groundwork for Meta's subsequent integrity and safety operations.

b. Research-Driven Technology

123. From a technological standpoint, pushing the state of the art in service integrity technologies is paramount. Not all services are equivalently engaged in the development of integrity systems and technologies. Some, but not all, services conduct state of the art research and publish scholarly work in various academic areas. Meta has been at the forefront for well over a decade in devoting substantial resources, efforts, and hiring of world known talent to study and address service integrity. Over a decade ago, for example, Meta instituted the Central Applied Science team ("CAS," and formerly known as the "Core Data Science" team).[172]

124. Meta has employed established academics to lead such research teams. For example, Dr. Lada Adamic, who until 2013 was a professor of computer science at the University of Michigan, served initially as manager then director of the CAS team from 2013 to 2020. Other academics have joined CAS, including Dr. Pablo Barberá (formerly a professor of political science at USC), Dr. Winter Mason (formerly a professor at Stevens Institute of Technology), and others.

125. The transparent publication and open documentation of service integrity protocols, technologies, and capabilities is another key element to advance public discourse and expert opinions. In this regard, not all social media service providers have engaged to the same extent with the public. Meta has a history of publishing highly impactful work documenting the internal

[172] *See* Meta. (2025). *Academic Papers Citations*, Central Applied Science. Retrieved June 10, 2025, from https://scontent-iad3-2.xx.fbcdn.net/v/t39.2365-6/475662651_598728696446916_8597606275327940151_n.pdf?_nc_cat=105&ccb=1-7&_nc_sid=e280be&_nc_ohc=NACZKDy1AAIQ7kNvwGbeWiG&_nc_oc=AdkKRBxzatks2X30fVU4C2t8ve3ZB-Jyiqv01DKJi-9HImDZfNJObcRkVwFc1Wv697Y&_nc_zt=14&_nc_ht=scontent-iad3-2.xx&_nc_gid=I6cirbBqvMSs0J0F3n4pQw&oh=00_AfPQKLbjafpiQFC135f8DeGHa3amSobO1kY6gcWC3TF_cQ&oe=686174C4 (showing scholarship published by CAS as early as 2009).

policies and systems that they design and deploy in production systems.

126. For example, a Facebook team led by Alon Halevy and collaborators, recently published a paper documenting the pipeline for enforcing integrity based on the guidelines developed at Meta. The paper appeared in the Communication of the ACM, the flagship computing magazine with a circulation of over 100,000 monthly readers.[173]

127. Meta has been also providing support to the research community by starting initiatives to share (in a privacy preserving and ethical way), datasets to enable studies on ads targeting, content sharing, and engagement on Facebook.[174]

2. How Classifiers Work

128. Building on years of technological iteration and research, Meta's suite of classifiers today are trained to look for specific signals and make predictions as to whether or not certain signals are present in a piece of content (e.g., nudity is present).[175] When a classifier is created, Meta trains the machine learning model by introducing more content of a specific type to improve the model's detection accuracy.[176]

129. Teams of reviewers make the final call on classifier predictions, and classifiers learn from those decisions.[177] Meta measures recall during the pre- and post-deployment phases of classifier development to assess the technology's performance,[178] and continually refines the

---

[173] See *Halevy, Alon, et al. "Preserving integrity in online social networks." Communications of the ACM 65.2 (2022): 92-98.*

[174] Meta. (n.d.). *Researcher datasets*. Retrieved June 10, 2025, from https://fort.fb.com/researcher-datasets.

[175] Meta. (2024, Nov. 12). *How enforcement technology works*. Meta Transparency Center. https://transparency.meta.com/enforcement/detecting-violations/how-enforcement-technology-works/.

[176] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 5-7, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[177] Meta. (2024, Nov. 12). *How enforcement technology works*. Meta Transparency Center. https://transparency.meta.com/enforcement/detecting-violations/how-enforcement-technology-works/.

[178] Meta Defendants' First Supplemental Responses and Objections to Plaintiffs' Fifth Set of Interrogatories at 2-3, No. 4:22-MD-03047-YGR (N.D. Cal. Mar. 17, 2025).

models.[179]

130. A classifier assigns a "Confidence Level" ("CL") to the content it reviews.[180] A CL is a statistical value referring to the probability that a piece of content is policy-violating (*e.g.*, p-value = 0.95); the higher the CL value, the more likely it is that the content is policy-violating. Conversely, the lower the CL value, the less likely it is that the content is policy-violating.

131. CLs are measured against "threshold values" to determine what type of action, if any, to take on a piece of content.[181] If the CL exceeds the applicable threshold value, generally the content will be automatically removed. If the CL is below the removal threshold, the classifiers may nevertheless "soft action" the content or route it for human review, as described above in Part IV.C.[182] Soft actions may vary depending on the category of content and type of classifier.[183] If a piece of content is assigned a CL that exceeds a certain threshold value, then that content is deemed to violate Meta's content policy and is automatically removed.[184] If a piece of content is assigned a CL that is lower than the removal threshold value, but higher than the threshold value that would trigger a soft action, then that content is routed for soft action and possible human review for a final determination.[185] If the confidence level is relatively low, content is not sent for human

[179] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 7, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[180] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 7, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[181] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 7, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[182] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 7, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[183] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 7, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[184] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 7-8, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[185] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 8, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

review; it is downranked, age gated, and/or subject to some other form of soft action.[186]

132. Threshold values may vary by the type of policy-violating content (e.g., SSI, ED) and may also vary across Facebook and Instagram. The threshold values are re-assessed both each time classifiers are updated and on an ongoing basis in order to maintain a consistent accuracy bar.

133. Where a piece of content violates multiple Community Standards, and any of the corresponding classifiers meet the threshold for deletion, the content will be removed.[187] Further, multiple classifiers can trigger soft actions at the same time, leading to even stronger demotion.[188] Further, I understand there are certain "multi-class" classifiers that can identify overlapping policy violations.[189]

134. The AI detection system runs proactively and automatically, with technology working behind the scenes to remove prohibited content, typically before any user sees it.[190] Meta also continuously trains its artificial intelligence system for more accurate and nuanced detection of prohibited content.[191] One way Meta does so is by training classifiers to look at all of the components of substantially similar posts, separate irrelevant differences from substantive differences (e.g., the presence of a Web browser header in one image and its absence from a substantially similar image would constitute an irrelevant difference), and then make a prediction

[186] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 8, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[187] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 8, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[188] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 8, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[189] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 8, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[190] Meta. (n.d.). *Prioritizing content review*. Retrieved June 10, 2025, from https://transparency.fb.com/policies/improving/prioritizing-content-review/.

[191]Meta. (n.d.). *How our enforcement technology works*. Retrieved June 10, 2025, from https://transparency.fb.com/enforcement/detecting-violations/how-enforcement-technology-works/.

of the contents' violative nature.[192] Another way is by using a system that guides AI to cull training data directly from millions of current content on Meta's services.[193] While this system also works by reactively reviewing user reports, out of all policy-violating content removed on Facebook and Instagram, the AI system generally removes over 90% proactively rather than through user reporting.[194]

135. Generally, if the item is obviously actionable (i.e., meets the threshold for deletion or other soft action), the system is self-executing. Machine learning of this nature learns and improves over time, and while it will never be perfect, its improvement has been observed in the Community Standards and Enforcement Reports.[195]

136. In addition to the classifiers discussed above, Meta also uses a banking system whereby content that is flagged for human review is enqueued in the "Single Review Tool," or SRT, to assess individual content, as described below in Part V.G. Policy-violating content is uploaded to the Media Match Service ("MMS"), which identifies new and historical copies of the content across Facebook and Instagram (and Messenger), which is automatically reviewed by Meta's proactive AI system, comparing "new" content to the MMS bank.[196] Virtually every piece of content uploaded to Facebook and Instagram (and Messenger) is checked against the MMS bank.[197] Groups of content identified as copies can be auto-enforced en masse, or they can be batched out for human review.[198] This en masse matching, auto-enforcement, and review process helps improve capacity, as Meta can avoid rote review of duplicate pieces of content, and can

[192] *See* Meta. (n.d.). *How our enforcement technology works*. Retrieved June 10, 2025, from https://transparency.fb.com/enforcement/detecting-violations/how-enforcement-technology-works/.

[193] Meta. (n.d.). *Training AI to detect hate speech in the real world*. Retrieved June 10, 2025, from https://ai.meta.com/blog/training-ai-to-detect-hate-speech-in-the-real-world/.

[194] Antigone Davis Dep. Vol. 1 at 81:1-5, Mar. 4, 2025.

[195] META3047MDL-001-00000112.

[196] META3047MDL-050-00003108.

[197] *See* META3047MDL-050-00003108.

[198] META3047MDL-050-00003108.

quickly address violating recycled content multiple times. Meta has used banking since at least 2011.[199]

137. Further, Meta maintains a "blackhole" database containing URLs that cannot be posted on Meta's services.[200]

a. Content Moderation of Live (Real-Time) Content

138. Moderation of Live content (i.e., content streamed or broadcast in real time) on Facebook and Instagram also involves the use of classifiers. However, there is a relative lack of data associated with Live content (in contrast to other content) because Live is not as heavily utilized by users.[201] Enforcement on Live presents an additional challenge because the content is not static; classifiers have to be rerun every 30 seconds of Live video to reevaluate whether the video is violating.[202] In contrast, pre-recorded video uploaded to Meta's services is analyzed by a classifier just once.[203] Furthermore, because of the nature of Live content, which can involve severe, rare, and immediate problems, moderation must be done near-instantaneously.[204] Moderation of Live is also more complex because it requires its own custom classifiers.[205]

139. Live content classifiers are newer relative to their static-content counterparts, and they frequently result in content being enqueued for human review, except for a few high-severity

[199] META3047MDL-208-00050921 (linking an "overview of bank names since 2011").

[200] *See* META3047MDL-003-00170869; META3047MDL-003-00138204.

[201] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 10, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[202] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 10, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[203] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 10, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[204] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 10, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025); *see also infra* Appendix A.

[205] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 10, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

harms.[206] Generally, enqueuing happens at around a 60-65% confidence level, with autodeletion when classifiers are at approximately above 85-90%.[207]

b. Abusive Account Moderation

140. When an account has repeatedly posted prohibited content, Meta makes the account harder to find on its services and limits the recommendation of that account's posts. Pursuant to Meta's "strike policy," if a user has a certain number of qualifying content takedowns—typically within a year-long period—their account is disabled.[208] Meta's strike policy is flexible, however, in that it may require fewer strikes before a takedown is actioned depending on the violation type—indeed, such action may occur in as little as a single post of prohibited content depending on the circumstances.[209] For example, posting "very high severity+" content[210] typically results in automatic account disablement.[211] When certain accounts are removed, Meta works to ensure that the device ID that created that account is blocked from creating new accounts in the app,[212] particularly those posting "very high severity+" content.[213] Some things that are violations of Community Standards do not result in strikes (e.g., spam).[214]

141. Meta automatically disables accounts that exhibit certain signals it monitors for

---

[206] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 10, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[207] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 10-11, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[208] *See* META3047MDL-046-00005060; META3047MDL-003-00062744; META3047MDL-003-00038376.

[209] *See, e.g.*, META3047MDL-014-00240391.

[210] META3047MDL-111-00372054. "Very high severity+" content is that which is "very rare, but can cause tremendous harm to" Meta's community and its reputation, and relevant categories of content include Child Safety and Credible Intent of Suicide (a subset of SSI) content. *Id.*

[211] META3047MDL-014-00240391.

[212] META3047MDL-004-00017504.

[213] META3047MDL-208-00051129; *see* META3047MDL-014-00240391 (discussing blocking the device of a user allegedly engaged in child sextortion).

[214] META3047MDL-046-00005060.

suspicious behavior. Meta identified and removed more than 90,000 accounts from August 1, 2023 to December 31, 2023 alone as a result of this monitoring.[215]

142. Where accounts have bios that include certain prohibited terms, those accounts are removed. For instance, an account with a bio that contains ED promotion (e.g., contains keywords like "thinspo, bonespo, thinspiration, etc."[216]) will ordinarily be removed. [217]

143. Abusive accounts are more resource intensive for Meta to review than individual posts or comments because they can take almost twice the amount of time for humans to review.[218]

c. Content-Specific Classifiers

144. Classifiers run automatically and review content on Facebook and Instagram. A robust and reasonable content moderation system must have discernable categories of policy violating content in order to scale effectively, and Meta's content-specific classifiers, as described below, offer such a solution. Different classifiers review and detect different categories of policy-violating content. The following sections describe generally the classifiers used to detect, categorize, and potentially action content across each of the relevant policy categories.

d. Suicide and Self-Injury Classifiers

145. I understand that Meta first implemented a classifier to action SSI content at least as early as February 2019, and that Meta has continuously developed and refined multiple SSI classifiers since then.[219] These classifiers use "pattern-recognition signals, such as phrases and comments of concern, to identify possible distress."[220] But, detection of SSI content is incredibly complex. For example, while humans might recognize that "I have so much homework I want to

---

[215] META3047MDL-034-00086282.

[216] META3047MDL-177-00000003.

[217]META3047MDL-003-00041618.

[218] META3047MDL-003-00070636.

[219] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 11, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[220] Meta. (n.d.). *Suicide prevention*. Retrieved June 10, 2025, from https://about.meta.com/actions/safety/topics/wellbeing/suicideprevention.

kill myself" is not a genuine cry for help, a computer does not have the contextual understanding of human nuance to not tag this phrase as alarming.[221] Similarly, a picture of train tracks may have to do with travel (and therefore be innocuous content), or it may relate to SSI.[222] Humans may have greater capacity than AI systems to more quickly tell the difference between these images based on the context, and training AI to do so is challenging, particularly at scale.[223] To train a classifier to have this contextual understanding it needs to be fed both positive examples (what you want it to identify) and contrasting negative examples (what you do not want it to identify) so it can learn to distinguish the two.[224] This requires a huge swath of data. [225]

146. Meta uses its AI technology, machine learning, and image-based technology (including pattern-recognition signals, such as phrases and comments of concern) to proactively identify and take action on clearly violating SSI content. SSI classifiers are refined by providing positive (i.e., actual SSI content) and negative (i.e., not actual SSI content, such as "I have so much homework, I want to kill myself") examples, so it can learn to distinguish the two.[226]

147. Meta's SSI classifiers review not only the content itself but also other information, for example, in the case of a post in which a user is seriously at risk, the classifier may detect and evaluate comments such as "Tell me where you are" or "has anyone heard from [user]?"[227]

148. Training automated technology to identify and distinguish between SSI content that

---

[221] Facebook. (2018, September 10). *How Facebook AI helps suicide prevention*. https://about.fb.com/news/2018/09/inside-feed-suicide-prevention-and-ai/.

[222] META3047MDL-001-00000112.

[223] META3047MDL-001-00000112; *see supra* Part III.B.

[224] Meta. (2018, September). *How Facebook AI helps suicide prevention.* https://about.fb.com/news/2018/09/inside-feed-suicide-prevention-and-ai/.

[225] Meta. (2018, September). *How Facebook AI helps suicide prevention*. https://about.fb.com/news/2018/09/inside-feed-suicide-prevention-and-ai/.

[226]Meta. (2018, September). *How Facebook AI helps suicide prevention*. https://about.fb.com/news/2018/09/inside-feed-suicide-prevention-and-ai/.

[227] Meta. (2018, September). *How Facebook AI helps suicide prevention*. https://about.fb.com/news/2018/09/inside-feed-suicide-prevention-and-ai/.

is violative or non-violative is highly complex, particularly on Instagram, where much of the content is image-based and therefore the intended meaning substantially relies on subtle context.[228]

e. Eating Disorder Classifiers

149. I understand Meta launched a classifier to action ED content specifically at least as early as March 2021.[229] Further, I understand Meta's ED classifiers have been continuously refined since then. Prior to the launch of an ED-specific classifier, I understand ED content fell under the purview of SSI policies and enforcement technology but were ultimately delineated into a unique harm area because of the challenges unique to detecting and reviewing ED content. Because Meta's services see less ED content than SSI content, there are fewer datapoints for the AI system to learn from in the ED sphere than the SSI sphere. To combat this issue, Meta generates synthetic training data the classifiers can learn from.

150. ED classifiers consider the following content as policy-violating and subject to removal: "[c]ontent that focuses on depiction of ribs, collar bones, thigh gaps, hips, concave stomach, or protruding spine or scapula when shared together with terms associated with eating disorders."[230] Not all policy violating ED content is subject to removal, such as content that "depicts ribs, collar bones, thigh gaps, hips, concave stomach, or protruding spine or scapula in a recovery context."[231] Rather, content of that nature is placed behind a sensitivity screen. Generally, content promoting or encouraging EDs are removed, but users are allowed to post content that touches on their own experiences around self-image and body acceptance.[232]

---

[228] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 12, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025)

[229] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 15, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[230] Meta. (n.d.). *Suicide and self-injury*. Retrieved June 10, 2025, from https://transparency.meta.com/policies/community-standards/suicide-self-injury/.

[231] Meta. (n.d.). *Suicide and self-injury*. Retrieved June 10, 2025, from https://transparency.meta.com/policies/community-standards/suicide-self-injury/.

[232] Meta. (n.d.). *Suicide and self-injury*. Retrieved June 10, 2025, from https://transparency.meta.com/policies/community-standards/suicide-self-injury/.

151. I understand that ED classifiers do not currently perform automatic deletions of content; rather, content is either queued for human review and subsequent removal or else soft-actioned.[233]

3. Bullying and Harassment Classifiers

152. I understand Meta first launched a classifier that actioned BH content at least as early as May 2018.[234] BH classifiers consider "severe" attacks on a public figure as policy-violating, but not in all cases content that merely degrades or shames a public figure.[235] Per Meta's policies, these classifiers consider public figures to be state and national level government officials, political candidates for those offices, people with over one million fans or followers on social media, and people who receive substantial news coverage. Further, BH classifiers consider content meant to degrade or shame any private individual as policy-violating.[236] Regardless of user status, Meta recognizes that bullying and harassment can have more of an emotional impact on minors, and extends heightened protection to anyone under the age 18.

153. Some content is removed for all users: repeatedly degrading or shaming contact, attacks based on the target's experience of sexual assault or domestic abuse, calls for SSI to a specific individual or group of individuals, derogatory terms, claims that a violent tragedy did not occur, claims that individuals are lying about being a victim of a violent tragedy, threats to release private information, statements to engage in a sexual activity, severe sexualized commentary, derogatory sexualized photoshop, calls for bullying or harassment of people, and degrading people who are depicted vomiting, defecating, urinating, or menstruating.

---

[233] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 16, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[234] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 17, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[235] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 18, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[236] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 18, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

154. In addition to Meta's tools to take action on a user's behalf, Meta also provides tools that help users control their own experiences on service, such as the ability to block certain key words from being posted in the comments to their content.

4. Graphic & Violent Content Classifiers

155. I understand Meta first launched a classifier that actioned VG content at least as early as December 2016.[237] VG classifiers are intended to detect graphic and violent content on Facebook and Instagram that may violate Meta's Policies.[238] VG classifiers consider, for example, photos and videos of wounded or dead people if they show dismemberment, visible innards, charred or burning people, victims of cannibalism, and/or throat slitting, as policy-violating.[239]

156. I understand that to protect users, and consistent with Meta's Community Standards, Meta employs a VG classifier to autodelete the most graphic and policy-violating content.[240] Other VG classifiers add a "disturbing" label to less graphic content to warn users that the content may be sensitive or disturbing and allow users to disengage before viewing the content.[241] When content is marked as disturbing by a VG classifier, the content is routed for human review and potential deletion. Conversely, Meta allows a subset of less graphic content to remain when users share content to shed light on or condemn acts.[242] Nevertheless, even for this less graphic content, various VG classifiers "age gate" (i.e., age restrict) content from youth

[237] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 23, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[238] Meta. (n.d.). *Violent & graphic content*. Retrieved May 4, 2025, from https://transparency.meta.com/policies/community-standards/violent-graphic-content/.

[239] Meta. (n.d.). *Violent & graphic content*. Retrieved May 4, 2025, from https://transparency.meta.com/policies/community-standards/violent-graphic-content/.

[240] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 23, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[241] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 23, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[242] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 23, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

users.[243]

5. Adult Nudity & Sexual Activity Classifiers

157. I understand Meta first launched a classifier to detect and action ANSA content at least as early as January 2016.[244] This group of classifiers detect images and videos of adult nudity and sexual activity on Facebook and Instagram that may violate policy.[245]

158. I understand adult nudity classifiers consider the following adult nudity content as policy-violating: "visible genitalia"; "visible anuses and/or fully nude close-ups of buttocks"; "uncovered female nipples"; "explicit sexual activity and stimulation"; "implicit sexual activity and stimulation"; "erections"; "presence of by-products of sexual activity"; "sex toys placed upon or inserted into mouth"; "stimulation of visible human nipples"; "squeezing female breasts"; "imagery depicting fetish that involves [ ] acts that are likely to lead to the death of a person or animal," "dismemberment," "cannibalism," "feces, urine, spit, snot, menstruation or vomit," "bestiality," and "incest"; "digital imagery of adult sexual activity"; and "extended audio of sexual activity."[246]

6. Child Safety Classifiers

159. Meta first launched a classifier to action "child safety" ("CS") content at least as early as 2018.[247] CS classifiers proactively identify and action content and behaviors that may violate Meta's Community Standards on Child Sexual Exploitation, Abuse, and Nudity.[248] (I

[243] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 23, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[244] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 25, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[245] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 25, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[246] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 25, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[247] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 27, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[248] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 27, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

understand that other experts are opining on Meta's child safety efforts more broadly, including Meta's work to detect, remove, and report child sexual abuse material, or CSAM, which are outside the scope of this report.)

160. Child Safety classifiers consider the following content as violating and subject to removal, among others: content, activity, or interactions that threaten, depict, praise, support, provide instructions for, make statements of intent, admit participation in, or share links of the sexual exploitation of children; content that solicits sexual content or activity depicting or involving children, nude imagery of real or non-real children, and/or sexualized imagery of real or non-real children; content that solicits sexual encounters with children; content that constitutes or facilitates inappropriate interactions with children; content that attempts to exploit real children by coercing money, favors, or intimate imagery with threats to expose intimate imagery or information or by sharing, threatening, or stating an intent to share private sexual conversations or intimate imagery; content that sexualizes real or non-real children; Groups, Pages, and profiles dedicated to sexualizing real or non-real children; content that depicts real or non-real child nudity; videos or photos that depict real or non-real non-sexual child abuse regardless of sharing intent; or content that praises, supports, promotes, advocates for, provides instructions for or encourages participation in non-sexual child abuse.[249]

161. Further, I understand that when users report potential CS content or when CS classifiers are triggered, the content is enqueued for human review, and Meta's review prioritization system helps to ensure that the potential CS content most likely to violate policy moves to the top of the review queue.[250] I understand that this process is in addition to Meta's other processes to detect, remove, and report CSAM, which are outside the scope of this report and subject to the reports of other experts.

[249] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 27-28, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[250] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 28, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

a. Borderline Classifiers

162. Meta employs advanced AI and machine learning tools to detect and monitor borderline content, which is, again, "content on Facebook and Instagram that is not strictly policy implicating but nonetheless considered potentially problematic or low quality."[251] These systems analyze patterns and signals that indicate content may be approaching the threshold of acceptability. Borderline classifiers complement Meta's content-specific classifiers as described above, balancing the technical benefit of discernable categories of policy violating content against the ambitious goal of reducing exposure to content that is not clearly negative for all users.

163. A 2018 study published in *New Media & Society* showed how users actively discuss strategies and successful approaches to circumventing Facebook content moderation guardrails, often downplaying the significance of their intended actions.[252] Meta's enforcement approach generally involves reducing exposure to borderline content through soft actions rather than removal, and some of those enforcement strategies are described below.

164. Due to the dynamic nature of these nuanced policies, continuous updates and training of these AI models are necessary to keep up with evolving content trends and tactics used to circumvent detection: approaches addressing challenges like the constantly changing violation trends, lack of precision for specific violations, and limited exploration of new violation types. Meta's enforcement approach generally involves reducing exposure to borderline content through soft actions rather than removal, and some of those enforcement strategies are described below.

b. Content Downranking

165. One of Meta's primary strategies for handling borderline content is to reduce its visibility. This can include "downranking," i.e., demoting such content in news feeds, search

[251] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 30, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025). I understand that there are no "borderline" content detection classifiers in the CS content category, *id.* at 28, as Meta intends its policies to address CS content as directly as possible.

[252] Gerrard, Y. (2018). *Beyond the hashtag: Circumventing content moderation on social media.* New Media & Society, 20(12), 4492-4511.

results, and recommendations to limit its reach without outright removal. By making borderline content less prominent, Meta aims to reduce its potential impact while avoiding direct censorship.

166. A recent academic study, published by researchers at the Swiss' EPFL in collaboration with a Meta scientist, showed that automated content moderation (comment deletion) decreased subsequent rule-breaking behavior.[253] Furthermore, the study found that "the effect of deletion on the affected user's subsequent rule-breaking behavior was longer-lived than its effect on reducing commenting in general, suggesting that users were deterred from rule-breaking but not from commenting." I find Meta's automated approach to reduce the visibility of certain borderline content quite reasonable and robust, as supported by the literature, as a means to decrease subsequent posting of such content without deterring users from expressing themselves.

c. User Warning Screens and Labels

167. Providing users with context through warning screens and labels, informing users of the nature of the content that a user must acknowledge before they can view it, can also help to mitigate the effects of borderline content, to alert users to potentially sensitive content, and to allow users to disengage with the post should they choose. For example, posts related to suicide or suicide attempts that are not policy-violating receive warning screens, as do violent and graphic posts, some forms of nudity, and posts raising awareness on bullying and harassment.[254] These labels and warning screens empower users to make informed decisions about the content they engage with and share.

B. Additional Automated Tools to Prevent Exposure to Harmful Content

168. The "Hidden Words" tool allows a user to set up his or her account such that posts

[253] Horta Ribeiro, M., Cheng, J., & West, R. (2023, April). Automated content moderation increases adherence to community guidelines. In Proceedings of the ACM web conference 2023 (pp. 2666-2676).

[254] Meta. (2025, Mar. 13). *Providing context on sensitive or misleading content*. Meta Transparency Center. https://transparency.meta.com/enforcement/taking-action/context-on-sensitive-misleading-content/ (discussing examples on Facebook); Meta. (n.d.). *Suicide and self-injury*. Retrieved June 10, 2025, from https://transparency.meta.com/policies/community-standards/suicide-self-injury/ (discussing similar warning screens on Instagram).

that contain certain words can be hidden from his or her feed on Instagram. The same feature applies to direct messages as well; users can turn on this feature to automatically filter direct message requests containing offensive words, phrases, and emojis,[255] placing them in a folder separate from innocuous messages so that users are shielded from negative messages unless they opt in.[256] Once a user opens the folder with the filtered messages, the message text is still automatically blurred until the user taps to uncover it.[257] The program was launched in April 2021.[258] Instagram worked with outside experts to develop a preset list of terms that will be filtered from DMs, and the user can also create a custom list of keywords that will be filtered out.[259]

169. "On-Device Nudity Control" ("ODNC") is a tool that detects unwanted nude content received in Instagram direct messages and automatically blurs it out. ODNC signals to the user that an image may contain nudity, and gives them the option to unblur the media, block the message, or get further help.[260] It is controlled by the user, and does not automatically report images back to Meta or law enforcement authorities.[261] It is on by default for users under 18

[255] Instagram. (n.d.). Hide comments or message requests you don't want to see on Instagram. Retrieved June 10, 2025, from https://help.instagram.com/700284123459336.

[256] META3047MDL-003-00001792; META3047MDL-053-00012559.

[257] Instagram. (n.d.). *Introducing new tools to protect our community from abuse*. Retrieved June 10, 2025, from https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse.

[258] Meta. (n.d.). *Our tools, features, and resources to help support teens and parents*. Retrieved June 10, 2025, from https://www.meta.com/help/policies/safety/tools-support-teens-parents/.

[259] Instagram. (n.d.). *Introducing new tools to protect our community from abuse*. Retrieved June 10, 2025, from https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse.

[260] Meta. (2024, April). *New tools to help protect against sextortion and intimate image abuse*. https://about.fb.com/news/2024/04/new-tools-to-help-protect-against-sextortion-and-intimate-image-abuse/.

[261] Meta. (2024, April). *New tools to help protect against sextortion and intimate image abuse*. https://about.fb.com/news/2024/04/new-tools-to-help-protect-against-sextortion-and-intimate-image-abuse/.

globally, and opt-in for everyone else.[262]

170. "Comment Filters/Comment Cover" are tools which block toxic, offensive, and divisive comments automatically.[263] The feature was first launched on Instagram in June 2017. It was expanded in May 2018 to include bullying and offensive comments, automatically hiding toxic comments, comments attacking a person's appearance or character.[264] Upon its implementation, Meta saw an 8% drop in bullying comment reports.[265] However, this feature experienced some shortcomings, such as difficulty recognizing foreign characters.[266] The comment filter is on by default, but users can turn it off in the Comment Controls center in the Instagram app. Additionally, there is a manual comment filter that enables the user to set their own words that trigger the filter.[267] Meta uses "fuzzy matching" to catch variations of prohibited words.[268]

171. "Sensitive Content Control" is a tool unique to Instagram which allows users to decide how much sensitive content shows up in Instagram's Explore tab, was launched in July

---

[262] Meta. (2024, April). *New tools to help protect against sextortion and intimate image abuse*. from https://about.fb.com/news/2024/04/new-tools-to-help-protect-against-sextortion-and-intimate-image-abuse/.

[263] Instagram. (2017, June 29). *Keeping Instagram a safe place for self-expression*. https://about.instagram.com/blog/announcements/keeping-instagram-a-safe-place-for-self-expression.

[264] Instagram. (n.d.). *Bully Filter and Kindness Prom to protect Instagram community*. Retrieved June 10, 2025, from https://about.instagram.com/blog/announcements/bully-filter-and-kindness-prom-to-protect-our-community; Meta. (n.d.). *Our tools, features, and resources to help support teens and parents*. Retrieved June 10, 2025, from https://www.meta.com/help/policies/safety/tools-support-teens-parents/.

[265] META3047MDL-003-00139760.

[266] META3047MDL-003-00058255.

[267] Instagram. (2017, June 29). *Keeping Instagram a safe place for self-expression*. Retrieved from https://about.instagram.com/blog/announcements/keeping-instagram-a-safe-place-for-self-expression.

[268] For instance, if "taco" is prohibited, through fuzzy matching, taaco, ta-co, and other variants would be caught as well. META3047MDL-003-00046768.

2021.[269] Sensitive content covers posts that are not against Meta's policies but may be upsetting to certain users (e.g., sexually suggestive or violent content).[270] Sensitive Content Control gives users control over the strictness of the filter with three options: More (allows users to see more sensitive content), Standard (default state), and Less (restricts more sensitive content).[271] Users under the age of 18 cannot opt into More sensitive content, as only the Standard and Less options are available.[272]

1. Public Efforts to Improve Content Moderation Technology

172. In recognition of the challenges of detecting harmful content, Meta open-sources some of its technology to make it available for others to use and improve.[273] For other content moderators who use their own content-matching technology, hash-sharing enables those systems to share digital fingerprints of problematic content to make detection easier.[274] Meta has also launched open competitions, in partnership with Microsoft, the Partnership on AI, academics from several universities, Getty Images, and Driven data, to attract talent to solve difficulties around

---

[269] Instagram. (n.d.). *Introducing new tools to protect our community from abuse*. Retrieved June 10, 2025, from https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse. As of a year after implementing, 99% of users who were defaulted in remain with this setting on. Meta. (2024, January). *Our work to help provide young people with safe, positive experiences*. Retrieved June 10, 2025, from https://about.fb.com/news/2024/01/our-work-to-help-provide-young-people-with-safe-positive-experiences/.

[270] Instagram. (n.d.). *Introducing new tools to protect our community from abuse*. Retrieved June 10, 2025, from https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse.

[271] Instagram. (n.d.). *Introducing new tools to protect our community from abuse*. Retrieved June 10, 2025, from https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse.

[272] META3047MDL-014-00243437.

[273] *See, e.g.*, Meta. (2019, Sept. 5). *Creating a dataset and a challenge for deepfakes*. Meta Blog. https://ai.facebook.com/blog/deepfake-detection-challenge/; Meta. (2019, August). *Open-sourcing photo- and video-matching technology to make the internet safer*. https://about.fb.com/news/2019/08/open-source-photo-video-matching/.

[274] Facebook. (2019, August). *Open-sourcing photo- and video-matching technology to make the internet safer*. https://about.fb.com/news/2019/08/open-source-photo-video-matching/.

detecting deepfakes and hateful memes. Two competitions that I am aware of were called the "Deepfake Detection Challenge,"[275] and the "Hateful Memes Challenge."[276]

173. *Deepfake Detection Challenge*: The Deepfake Detection Challenge was intended to "produce technology that everyone can use to better detect when AI has been used to alter a video in order to mislead the viewer."[277] It includes monetary grants and awards to motivate the industry to innovate in this space, and is overseen by "the Partnership on AI's new Steering Committee on AI and Media Integrity," which is made up of a broad cross-sector coalition of organizations including Meta (then-Facebook), WITNESS, Microsoft, and others in civil society and the technology, media, and academic communities.[278] Meta invested heavily in this challenge (over $10 million), through which it has tried to create as realistic a dataset as possible in order to get the most useful results.[279]

174. *Hateful Meme Challenge*: Similarly, the Hateful Memes challenge was a first-of-its-kind $100,000 competition intended to accelerate research on the problem of detecting hate speech that combines both images and text.[280] It had more than 3,300 participants around the world, and the winning model achieved a 0.8450 AUC ROC, far exceeding the baseline model.[281]

---

[275] Meta. (2019, Sept. 5). *Creating a dataset and a challenge for deepfakes*. Meta Blog. https://ai.facebook.com/blog/deepfake-detection-challenge/.

[276] Meta. (2020, Dec. 11). *Hateful memes challenge winners*. Meta Blog. https://ai.facebook.com/blog/hateful-memes-challenge-winners/.

[277] Meta. (2019, Sept. 5). *Creating a dataset and a challenge for deepfakes*. Meta Blog. https://ai.facebook.com/blog/deepfake-detection-challenge/.

[278] Meta. (2019, Sept. 5). *Creating a dataset and a challenge for deepfakes*. Meta Blog. https://ai.facebook.com/blog/deepfake-detection-challenge/.

[279] Meta. (2019, Sept. 5). *Creating a dataset and a challenge for deepfakes*. Meta Blog. https://ai.facebook.com/blog/deepfake-detection-challenge/.

[280] Meta. (2020, Dec. 11). *Hateful memes challenge winners*. Meta Blog. https://ai.facebook.com/blog/hateful-memes-challenge-winners/.

[281] Meta. (2020, Dec. 11). *Hateful memes challenge winners*. Meta Blog. https://ai.facebook.com/blog/hateful-memes-challenge-winners/.

2. Other Efforts to Improve Content Moderation Technology

175. Models developed by Meta AI scientists, such as XLM-R,[282] leverage cross-lingual training to understand and moderate content across different languages, enabling them to apply knowledge gained from one language to another.

176. Meta develops and employs advanced AI transformer models, such as the RoBERTa[283] and XLM-R[284] architectures. Transformer models such as these are at the forefront of detecting complex and subtle forms of harmful content. These models rely on mechanisms to focus on relevant parts of the input data, allowing them to understand the context and nuances of language and imagery.

177. Another notable advancement is the Linformer architecture, developed by Meta's scientists, which enhances the efficiency of Transformer models, enabling the processing of longer text inputs with less computational overhead.[285]

C. Meta's Reporting System

178. While Meta's proactive mechanisms identify the bulk of the harmful content posted to Meta's services, given the immense volume and highly varied nature of the content on Meta's services, no technology could ensure that 100% of all policy-violating content is appropriately actioned. Thus, Meta has a "safety net" of reactive content moderation that can review and take

---

[282] *See* Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. *Unsupervised cross-lingual representation learning at scale*. arXiv preprint arXiv:1911.02116 (2019). Note that this paper has been cited over 7,000 times by researchers in the field of AI, demonstrating its significant impact.

[283] *See* Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[284] *See* Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). *Unsupervised cross-lingual representation learning at scale*. arXiv preprint arXiv:1911.02116

[285] *See* Wang, Sinong, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. *Linformer: Self-attention with linear complexity*. arXiv preprint arXiv:2006.04768 (2020). Note that this paper has been cited over a thousand times by researchers in the field of AI, substantiating its contribution and importance.

action on content that automated systems did not. This includes content that is reported by users. Instagram alone receives roughly 10 million reports every day.[286]

179. Adding to the challenge of scale, user reports often lack reliable indicators of the actual saliency of policy violating content. The user reporting feature has been known to be abused: Academic research has shown that reporting tools are often used adversarially rather than to flag legitimate violations.[287] For example, "mass reporting brigades" are common on Twitter/X[288] and Reddit,[289] as well as Meta's services.[290] Further, ideological weaponization of reporting systems disproportionately impacts marginalized users.[291] Services can struggle to identify malicious use of moderation tools, especially when decisions are automated or unverified.

180. In addition, a Cornell researcher recently postulated the notion that users have been known to make use of higher-severity issue tagging for lower-severity violations in their reports in hopes of expediting response times or engage in harassment.[292] Reports are therefore not always an accurate signal of a potential violation type, which Meta's systems take into account.

---

[286] META3047MDL-019-00093833.

[287] Myers West, S. (2018). *Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms*. New Media & Society, 20(11), 4366-4383; Gorwa, R., Binns, R., & Katzenbach, C. (2020). *Algorithmic content moderation: Technical and political challenges in the automation of platform governance*. Big Data & Society, 7(1), 2053951719897945.

[288] Zannettou, S., Caulfield, T., Setzer, W., Sirivianos, M., Stringhini, G., & Blackburn, J. (2019, June). *Who let the trolls out? towards understanding state-sponsored trolls*. In Proceedings of the 10th ACM Conference on Web Science (pp. 353-362).

[289] Jhaver, S., Appling, D. S., Gilbert, E., & Bruckman, A. (2019). "Did you suspect the post would be removed?" Understanding user reactions to content removals on Reddit. Proceedings of the ACM on human-computer interaction, 3(CSCW), 1-33.

[290] *See, e.g.*, Meta. (2021, December). Meta's adversarial threat report. Retrieved June 10, 2025, from https://about.fb.com/news/2021/12/metas-adversarial-threat-report/#:~:text=Mass%20Reporting%3A%20We%20will%20remove,Behavior%20policy%20against%20mass%20reporting.

[291] Marwick, A. E., & Caplan, R. (2018). *Drinking male tears: Language, the manosphere, and networked harassment.* Feminist Media Studies, 18(4), 543-559.

[292] Meisner, C. (2023). *The weaponization of platform governance: Mass reporting and algorithmic punishments in the creator economy*. Policy & Internet, 15(4), 466-477.

181. As noted elsewhere in this report, users are typically not exposed to very much prohibited content.[293] In the rare instance where potentially prohibited content is not filtered by the AI system, users are able to report such content through in-app reporting functions, which are available on all devices.[294]

182. Meta employs metrics to assess and continually improve action taken on user reports, such as Recall over All Reports ("RoAR"), which was replaced by Recall over Tagged Reports ("RoTR") in 2024.[295] RoAR asked what percentage of user reports on actually violating content did Meta enforce within N days.[296] RoTR, the evolution of the RoAR metric, measures the percentage of reported and confirmed "very high severity+" content that is appropriately actioned within 96 hours.[297]

D. Meta's Human Review Processes

183. Meta has thousands of reviewers globally who review potential violations of Facebook and Instagram policies, which is called the Global Operations team.[298] I will first provide an overview of the review tool, the review process, and escalation discuss how Meta undertakes this effort.

1. Overview

184. Meta's human reviewers use the "Single Review Tool" or SRT to assess individual

---

[293] *See infra* Part VI (discussing the prevalence rate in Meta's CSER report); Antigone Davis Dep. Vol. 1 at 81:1-5, Mar. 4, 2025.

[294] Instagram. (n.d.). *Report a post or profile on Instagram*. Instagram Help Center. Retrieved June 10, 2025, from https://help.instagram.com/192435014247952/; Meta. (2023, Oct. 18). *How technology detects violations*. Meta Transparency Center. https://transparency.meta.com/enforcement/detecting-violations/technology-detects-violations/; META3047MDL-001-00000112.

[295] META3047MDL-208-00061906.

[296] META3047MDL-208-00050989.

[297] META3047MDL-208-00061906; *see supra* footnote 210 (defining "very high severity+" content).

[298] Meta. (n.d.). *How review teams work*. Meta Transparency Center. Retrieved June 10, 2025, from https://transparency.meta.com/enforcement/detecting-violations/how-review-teams-work/.

content. What reviewers see depends on what is reported. For instance, where Instagram comments are reported, the reviewer sees the Instagram image, the caption, and the comment under review.[299] They then can determine whether the comment complies with the Community Standards and Policy Labs.[300] At times, the reviewer simply looks at the content to determine if it is violates Meta's policies, but at other times, a reviewer may be encouraged to look at the context of the content before reaching a decision.[301]

185. If a human reviewer needs help, they can escalate the questionable content to a subject matter expert (i.e., someone specialized in the review of a particular category of content) to decide.[302] Additionally, where a reviewer is concerned about an imminent credible threat of physical harm, the content is escalated to safety experts who determine the threat's credibility and whether it should be reported to emergency services.[303]

186. Meta conducts regular audits and quality checks to help ensure that reviewers are applying standards consistently and accurately.[304] These audits help identify areas for improvement in training and guidelines.

187. Meta also has a system in place to allow for two kinds of appeals—"actor appeals," where a user appeals the decision to take down something they themselves posted; and "reporter appeals" where the reporter asks Meta to "seriously, take a look" at something that was *not* removed.[305] Similar to first-level user reports, appeals are reviewed by Meta's automated systems for review prioritization.

---

[299] META3047MDL-003-00036694.

[300] META3047MDL-003-00036694.

[301] META3047MDL-003-00036694.

[302] META3047MDL-001-00000112 .

[303] META3047MDL-001-00000112 .

[304] META3047MDL-003-00001478; META3047MDL-050-00084511.

[305] META3047MDL-003-00144080. By at least Feb. 11, 2019, actor appeals were enabled for most problem types, and reporter appeals were enabled across the board.

2. In-House & Third-Party Human Reviewers

188. Meta has around 35,000 internal and third-party human reviewers. Internal reviewers are closely-integrated with Meta's overall content moderation strategy. This integration helps ensure that content moderation is consistent with company policies and that human reviewers are aligned with the company's mission and standards.

189. Internal reviewers can provide immediate feedback on policy effectiveness and suggest improvements. This direct line of communication helps in refining content policies and enforcement mechanisms, leading to more accurate enforcement outcomes.

190. Inviting third-party reviewers into these processes offers scalability, allowing Meta to handle large volumes of flagged content more efficiently. This is particularly important during peak times, such as during major events or crises when the volume of content can surge.

191. One of the main challenges with both internal and third-party review is the potential for variations in enforcement. Different reviewers may have different interpretations of policies, leading to inconsistencies. To mitigate this, Meta provides detailed guidelines and training on applying the Community Standards to third-party reviewers.[306]

192. If a human reviewer remains unsure whether a piece of content violates the Community Standards, they can escalate the questionable content to a content policy subject matter expert on the Global Operations team to help the reviewer come to a decision.[307] Additionally, where a reviewer is concerned about an imminent credible threat of physical harm, the content is escalated to experts who determine the threat's credibility and whether it should be reported to emergency services.[308]

193. The human review system also provides feedback to the AI system to further

[306] Meta. (n.d.). *Helping reviewers make the right calls*. Meta Transparency Center. Retrieved June 10, 2025, from https://transparency.meta.com/enforcement/detecting-violations/making-the-right-calls/.

[307] META3047MDL-001-00000112.

[308] META3047MDL-001-00000112.

improve it.[309] For example, every comment is human reviewed at least twice before it is fed into the AI system to ensure accuracy.[310] Importantly, when the human review team makes particularly difficult judgment calls, those decisions are fed to the AI system to further mimic them so that more prohibited content can be reviewed proactively.[311]

3. Prioritization of Content for Human Review

194. Given the volume of content on Meta's services,[312] human review on all content reported by users or detected by AI is impracticable. Instead, given the realistic limits of technology and human capacity that Meta faces (as would be the case for any social media service operating at Meta's scale), Meta triages the review of content by humans. The following describes the ranking and review process for organic content (i.e., non-ad content posted by users of Facebook and Instagram).

195. Before an item reaches human reviewers, Meta's technology ranks and prioritizes content so that its team of reviewers is presented with and can focus on the most important cases first.[313]

196. Once a piece of content is enqueued for human review, then that piece of content is assigned a predictive integrity value ("pIV"), which is used to determine which piece of content

[309] Meta. (2024, Nov. 12). *How Meta prioritizes content for review*. Meta Transparency Center. https://transparency.meta.com/policies/improving/prioritizing-content-review/.

[310] META3047MDL-003-00036694.

[311] Meta. (2024, Nov. 12). *How enforcement technology works*. Meta Transparency Center. https://transparency.meta.com/enforcement/detecting-violations/how-enforcement-technology-works/.

[312] *E.g.*, Smith, C. (2013, September 18). *Facebook users are uploading 350 million new photos each day*. Business Insider. https://www.businessinsider.com/facebook-350-million-photos-each-day-2013-9.

[313] Meta. (2022, Jan. 19). *How technology helps prioritize review*. Meta Transparency Center. https://transparency.meta.com/enforcement/detecting-violations/technology-helps-prioritize-review.

should be reviewed first by the human reviewers.[314] PIV is calculated based on factors including likelihood of violation, virality, and violation severity, as further explained below.[315]

197. At this review stage, I understand a range of signals further contribute to the prediction that a piece of content is policy-violating content for the purpose of prioritization, including Whole Post Integrity Embeddings ("WPIE"). WPIE is a holistic, multimodal, and robust classification layer—in addition to content-specific classifiers described above—that looks at a swath of contextual queues (e.g., image, text, comments) to predict whether a piece of content is policy violating.[316] WPIE assesses the basic details of the content being detected to determine what kind of violation may be at issue.[317] As I describe above,[318] multimodal solutions are quite resource-intensive and require advanced models and large sets of training data to develop and improve upon. At the scale that Meta processes user reports and escalations from first-level classifiers, I find a multimodal solution like that presented by WPIE to be a robust and more-than-reasonable measure toward the protection of users.

198. Meta's review prioritization efforts do not stop there. Meta's Multi-Armed Bandit ("MAB") system assesses the content's severity level and assigns a "predictive severity level" weight to the content based on the policy area it is most likely to potentially violate (i.e., if MAB suspects the content represents a very high severity problem it will get a higher severity weight,

[314] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 9, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[315] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 9-10, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[316] META3047MDL-014-00055649; META3047MDL-014-00073104; *see also* META3047MDL-014-00328102; META3047MDL-072-00001183; META3047MDL-031-00069986.

[317] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 9, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[318] *See* discussion *supra* Part III.B.3.

while if it is a high or medium severity the assigned weight will be lower).[319]

199. Other ranking systems evaluate content virality. The Virality, Estimations, Prediction, and Applications ("VESPA") system assesses the piece of content's virality by counting each time a user views it for a meaningful amount of time. VESPA assigns a "predicted virality" value to a content. "Predicted virality value" reflects how widely viewed a particular piece of content is expected to be.[320] The High Risk Early Review Operations ("HERO") system is an end-to-end system used for human review to remove high viral violating content, including content flagged by VESPA.[321]

200. Based on these evaluations of harm classification, severity, and virality, content is assigned a pIV to prioritize the piece of content for review.[322] A priority formula then orders content for review based on pIV.[323] The highest pIV tasks move to the top of the queue so that the most viral and potentially harmful content gets reviewed sooner.[324] The queue is designed to be dynamic, such that it continually brings the highest priority items to the top of the queue for review. Some content is always prioritized because it is "extremely high severity" content (e.g., content that may violate Meta's child safety policies), and even if they have very low virality, the severity rating will carry it to the top of the review queue.[325]

---

[319] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 9, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025). Meta does not assign "weight" to or otherwise "rank" classifiers. Rather, content is ranked for human review, according to its predicted severity. *Id.*

[320] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 9-10, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[321] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 9-10, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[322] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 10, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[323] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 10, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[324] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 10, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

[325] Meta Defendants' Supplemental Responses and Objections to Plaintiffs' First Interrogatory at 10, No. 4:22-MD-03047-YGR (N.D. Cal. Feb. 21, 2025).

VI. Effectiveness and Transparency of Meta's Content Moderation Systems

A. CSERs & the Prevalence Metric

201. Meta has published quarterly Community Standards Enforcement Reports ("CSERs") since 2018.[326] These reports measure prevalence (the estimated percentage of views that were of violating content when compared to all views of content on Facebook or Instagram).[327] The most recent CSER data can be found online.[328]

202. More specifically, prevalence of violating content is estimated using samples of views of content from across Facebook or Instagram.[329] It is calculated as the estimated number of views that showed violating content, divided by the estimated number of total content views on Facebook or Instagram.[330] So, for example, if the prevalence of some category of violating content was 0.05% to 0.06%, that would mean of every 10,000 content views, 5 to 6 on average were of content that violated Meta's standards for adult nudity and sexual activity.[331]

B. Reasonableness of CSERs for Evaluating the Effectiveness of Content Moderation System

203. Meta's view-based "prevalence" metric is an appropriate measure for evaluating user exposure to harmful content. From a risk-assessment standpoint, what matters is not merely

---

[326] *See* Rosen, G. (2018, May 15). *Facebook publishes enforcement numbers for the first time*. Meta Newsroom. https://about.fb.com/news/2018/05/enforcement-numbers/; (Harwell & Timberg, 2020).

[327] Meta. (2025, Mar. 6). *Prevalence*. Meta Transparency Center. https://transparency.meta.com/policies/improving/prevalence-metric/.

[328] *See* Meta. (n.d.). *Community standards enforcement report*. Meta Transparency Center. Retrieved June 13, 2025, from https://transparency.meta.com/reports/community-standards-enforcement/.

[329] Meta. (2025, Mar. 6). *Prevalence*. Meta Transparency Center. https://transparency.meta.com/policies/improving/prevalence-metric/.

[330] Meta. (2025, Mar. 6). *Prevalence*. Meta Transparency Center. https://transparency.meta.com/policies/improving/prevalence-metric/.

[331] Meta. (2025, Mar. 6). *Prevalence*. Meta Transparency Center. https://transparency.meta.com/policies/improving/prevalence-metric/.

how much violating content exists on the service, but how much of it is seen. A piece of violating content that is uploaded but quickly removed without exposure poses virtually no harm. Meta's focus on exposure to policy violating content reflects a scientifically grounded approach, consistent with risk-based frameworks used in safety engineering and public health surveillance.[332]

204. The above-described sampling methodology is statistically rigorous and representative, and permits inference about service-wide behavior. This approach reflects current best practices in algorithmic accountability and harm prioritization.

205. Meta also reports several related metrics in its CSERs, including: the number of content items actioned; the "proactive rate" (i.e., the percentage of all content or accounts acted on that Meta found and actioned before users reported them to Meta); the number of appeals; and the number of restored content items. These multiple indicators enable evaluation not only of coverage (recall) but also of precision, fairness, and responsiveness. This multi-metric framework is consistent with the kinds of internal validation and external audit processes used in mature, safety-critical systems.

206. While no measurement system is immune to limitations, Meta's CSER framework is transparent, statistically grounded, and continuously refined. Meta discloses methodology, publishes quarterly data, and permits independent review. In my opinion, this level of transparency and methodological integrity substantially exceeds industry norms.

207. Lastly, it is worth noting that the prevalence metric is well supported in the academic literature.[333] Leading scholars in the fields of algorithm auditing and service

[332] Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and discrimination: converting critical concerns into productive inquiry, 22(2014), 4349-4357; Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018, April). 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In Proceedings of the 2018 Chi conference on human factors in computing systems (pp. 1-14).

[333] Yale Law School. (2019, May 23). *Facebook data transparency advisory group releases final report*. https://law.yale.edu/yls-today/news/facebook-data-transparency-advisory-group-releases-final-report.

accountability have advocated for view-based exposure metrics as superior indicators of user harm compared to raw content counts.[334] The CSER framework aligns closely with these recommendations and reflects an empirically responsible posture toward reduction of policy-violating content and transparency:

208. First, CSER is based on prevalence, i.e., the estimated proportion of total content views that contained policy-violating material, a measure rooted in audience exposure, not just policy enforcement.

209. Second, CSER provides category-specific metrics (e.g., hate speech, graphic violence, adult nudity) and includes confidence intervals, reflecting best practices in measurement transparency and statistical disclosure.

210. Third, Meta applies probability-based stratified sampling and then applies manual review procedures, ensuring that prevalence estimates in CSER are both replicable and independently auditable, two key standards recommended by both the Facebook Data Transparency Advisory Group and independent researchers.

211. Finally, Meta regularly updates its methodology and discloses changes to its classifiers and enforcement systems over time, which mirrors the kind of iterative evaluation frameworks recommended in transparency and auditing literature.

C. Most-Recent CSER Report

212. Meta's most recent CSER report was published in February 2025.[335] That report included measurements on prevalence, content actioned, proactive rate, appealed content, and restored content, spanning the policy areas of violence and criminal behavior, safety, objectionable

---

[334] Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 2053951719897945; Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A., & Meira Jr, W. (2020, January). Auditing radicalization pathways on YouTube. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 131-141).

[335] *See* Meta. (n.d.). *Community standards enforcement report*. Meta Transparency Center. Retrieved June 13, 2025, from https://transparency.meta.com/reports/community-standards-enforcement/.

content, and integrity and authenticity. Meta presently defines these metrics as follows:

213. Prevalence: estimated percentage of views that were of violating content (defined in more detail above).[336]

214. Content actioned: the number of pieces of content (such as posts, photos, videos or comments) or accounts taken action on for going against standards; shows the scale of enforcement activity.[337]

215. Proactive rate: the percentage of all content or accounts acted on that Meta found and actioned *before* users reported them to Meta; used as an indicator of how effectively Meta detects violations.[338]

216. Appealed content: the number of pieces of content (such as posts, photos, videos or comments) that people appeal after Meta takes action on for going against policies.[339]

217. Restored content: the number of pieces of content (such as posts, photos, videos or comments) restored after Meta originally took action on them.[340]

218. As of the Q1 2025 report, CSER published prevalence metrics from Instagram and Facebook for the following relevant categories of content:

Table 6. Adult Nudity & Sexual Content CSER

Metric	Instagram	Facebook
Prevalence	Lower bound and upper bound: 0.04%	Lower bound: 0.05%; Upper bound: 0.06%
Content Actioned	9.8M	64.7M
Content Appealed	1.6M	2.5M

[336] Meta. (2025, Mar. 6). *Prevalence*. Meta Transparency Center. https://transparency.meta.com/policies/improving/prevalence-metric/.
[337] Meta. (2023, Nov. 7). *Content actioned*. Meta Transparency Center. https://transparency.meta.com/policies/improving/content-actioned-metric/.
[338] Meta. (2023, Feb. 22). *Proactive rate*. Meta Transparency Center. https://transparency.meta.com/policies/improving/proactive-rate-metric/.

[339] Meta. (2022, Nov. 18). *Appealed content*. Meta Transparency Center. https://transparency.meta.com/policies/improving/appealed-content-metric/.

[340] Meta (2022, Oct. 4). *Restored content*. Meta Transparency Center. https://transparency.meta.com/policies/improving/restored-content-metric/.

Metric	Instagram	Facebook
Proactive Rate	97.7%	94.9%
Restored Content	825.1K with appeal; 97.8K without appeal	1.1M with appeal; 191.2K without appeal

Table 7. Bullying & Harassment CSER

Metric	Instagram	Facebook
Prevalence	Lower bound: 0.05%; Upper bound: 0.06%	Lower bound: 0.07%; Upper bound: 0.08%
Content Actioned	5.2M	5.1M
Content Appealed	778.3K	904K
Proactive Rate	88%%	73.3%
Restored Content	196.9K with appeal; 79.4K without appeal	143.6K with appeal; 23.4K without appeal

Table 8. Child Endangerment (nudity and physical abuse) CSER

Metric	Instagram	Facebook
Prevalence	N/A	N/A
Content Actioned	616K	1.5M
Content Appealed	58.8K	163K
Proactive Rate	98%	97.2%
Restored Content	23.5K with appeal; 7.2K without appeal	32KK with appeal; 42.7K without appeal

Table 9. Child Endangerment (sexual exploitation) CSER

Metric	Instagram	Facebook
Prevalence	N/A	N/A
Content Actioned	1.5M	4.6M
Content Appealed	93.2K	398.1K
Proactive Rate	94.5%	95.8%
Restored Content	28.9K with appeal; 32.5K without appeal	127K with appeal; 287.4K without appeal

Table 10. Suicide & Self-Injury CSER

Metric	Instagram	Facebook
Prevalence	Upper bound: 0.05%	Upper bound: 0.05%
Content Actioned	9.9M	6.8M
Content Appealed	118.3K	231.6K
Proactive Rate	99.10%	98.9%
Restored Content	24.7K with appeal; 1.2K without appeal	41.3K with appeal; 1.5K without appeal

Table 11. Violent & Graphic Content CSER

Metric	Instagram	Facebook
Prevalence	Lower bound and Upper bound: 0.06%	Lower bound: 0.09%; Upper bound: 0.09%
Content Actioned	4.9M	10.7M
Content Appealed	26K	64.2K
Proactive Rate	97.5%	98.0%
Restored Content	10.3K with appeal; 31.4K without appeal	12.3K with appeal; 12.5K without appeal

219. Meta's most recent Community Standards Enforcement Report discloses several quantitative metrics that collectively provide a robust and multi-dimensional picture of the service's content moderation efficacy.

220. Proactive Rate indicates the percentage of violating content detected and actioned before users report it. High proactive rates, such as the 98.9% proactive detection of suicide and self-injury content on both Facebook and Instagram, demonstrate the success of Meta's AI-driven classifier systems in identifying policy-violating content without relying on user intervention.

221. The fact that only a relatively small portion of actioned content is appealed, and that an even smaller subset of those appeals result in restoration after human review, suggests that the enforcement system is both accurate and subject to meaningful user recourse.

In conclusion, while no moderation system is perfect, Meta's publication of granular, view-based prevalence data, superior proactive detection rates, and classifier audit disclosures meet or surpass transparency and enforcement norms across the social media services industry. These features lend empirical and comparative support to the reasonableness and good-faith design of its moderation framework.

D. Expert Assessment of the Algorithmic Stress Test of Instagram's Reels Surface

222. In this section, I evaluate the purported testing protocol of Instagram accounts that Mr. Arturo Bejar, a former employee and contingent worker at Meta who worked on user safety and security, claims he conducted from late 2023 to early 2024 "to see what the experience was of

a teenager on Instagram."[341] To evaluate Mr. Bejar's testing, I have reviewed relevant portions of his deposition transcripts and the accompanying exhibits, with particular attention to Exhibit 12.[342] My opinion of Mr. Bejar's testing protocol can be summarized as follows:

223. *First*, although Mr. Bejar does not characterize his testing protocol as such, functionally, Mr. Bejar appears to have conducted a truncated form of algorithmic stress test on Instagram's Reels surface.

224. *Second*, as an algorithmic stress test, Mr. Bejar's protocol and methodology suffer from fatal deficiencies that render the results scientifically unreliable.

225. I reach these conclusions based on the following evaluation.

1. Mr. Bejar Conducted a Truncated Algorithmic Stress Test

226. Mr. Bejar claims he evaluated "a few test accounts" "to see what the experience was of a teenager on Instagram."[343] Though Mr. Bejar does not identify his evaluation as such, Mr. Bejar's deposition testimony and accompanying exhibits indicate that he attempted an algorithmic stress test of Instagram's Reels content recommendation algorithm.

227. Algorithmic stress tests are forms of audits that are performed on algorithmic systems to test and replicate the behavior of the system under certain controlled conditions. When properly conducted, stress tests are part of an established toolkit of algorithmic auditing techniques that are acknowledged as valid inquiry tools by the scientific community.[344]

---

[341] Arturo Bejar Dep. Vol. 1 at 197:1-4, Apr. 7, 2025.

[342] *See* Arturo Bejar Dep. Vol. 1, Apr. 7, 2025; Arturo Bejar Dep. Vol. 2, Apr. 8, 2025; Arturo Bejar Dep. Vol. 3, Apr. 9, 2025.

[343] Arturo Bejar Dep. Vol. 1 at 197:1-4, Apr. 7, 2025.

[344] See the vast body of literature that rigorously outlines protocols for algorithmic audits, including:

(1) Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). *Auditing algorithms: Research methods for detecting discrimination on internet platforms*. Data and discrimination: converting critical concerns into productive inquiry, 22(2014), 4349-4357.

228. Though his methods were scientifically flawed, Mr. Bejar appears to have attempted an assessment of the behavior of Instagram's Reels algorithm under particular conditions, which suggests he attempted to conduct an algorithmic stress test. Using an Apple iPhone 11 and an iPad at indeterminate times, Mr. Bejar created a small yet indeterminate number of synthetic Instagram accounts through unique iCloud email addresses. Four accounts purported to represent minor female users, an indeterminate number of accounts purported to represent minor male users, and one account purported to represent a 23- or 25-year-old of indeterminate sex and gender, as a purported control.[345] Mr. Bejar manually steered some of these accounts toward content categories on Reels he deemed harmful or inappropriate (e.g., sexually suggestive material, self-harm–related content), or media involving body image issues or minors in potentially suggestive contexts. These interactions were recorded via screenshots and screen-capture videos, which were then presented as purported evidence that Instagram's recommender system promotes or amplifies harmful content to underage users.

229. While the concept of a "stress test" has a recognized place in both academic and engineering practice, Mr. Bejar's implementation of this method fails to conform to any scientifically accepted protocol for conducting such audits. In the context of algorithmic systems, a legitimate stress test is a *controlled, repeatable, and statistically grounded procedure* that deliberately subjects a system to adverse or edge-case inputs to evaluate its robustness, performance limits, or failure modes. These tests must be carefully designed to isolate causal relationships between input behaviors and system responses and must be accompanied by appropriate control conditions, transparency of design, and rigorous statistical evaluation. Mr.

(2) Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). *Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing*. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 33-44).

[345] BEJAR0002658 (indicating the account purported to represent a 23-year-old); Arturo Bejar Dep. Vol. 2 at 424:16-19, Apr. 8, 2025 (indicating the account purported to represent a 25-year-old).

Bejar's so-called test contains none of these critical elements.

230. If the goal is to evaluate how recommendation systems operate under stress or boundary conditions, particularly in safeguarding minors, a scientifically valid audit would ordinarily follow a transparent, causally informative experimental design. This entails several methodological components:

231. ***Clearly Defined Harm Categories***: First, researchers must pre-register[346] operational definitions for the types of harmful content being audited (e.g., pro-eating disorder imagery, sexually suggestive material involving minors, graphic self-harm). These definitions should be aligned with platform policies and third-party standards (e.g., WHO, NSPCC) to ensure validity.

232. ***Standardized Behavioral Profiles***: Test accounts must simulate real user behaviors (e.g., pausing on sensitive content, following triggering accounts, liking specific hashtags) using pre-scripted behavioral templates. These should be uniformly applied across conditions to reduce confounds.

233. ***Randomized Assignment and Control Groups***: A core principle of causal inference is randomization. Accounts should be randomly assigned to different behavioral archetypes (e.g., neutral, vulnerable, exploratory), with control accounts that exhibit neutral or passive browsing. This creates counterfactual conditions that allow for attributing observed effects to the manipulated behaviors.

234. ***Temporal Controls***: Since recommendation algorithms evolve over time,

[346] In simplistic terms, *pre-registration* means that before starting a study or audit, the researcher(s) would state and pre-register (write down in advance) what phenomenon or effect they're going to look for, how they'll look for it, and how they'll measure it. In the specific case, this would include clearly defining what counts as harmful content (like what exactly qualifies as pro-eating disorder imagery or sexually suggestive material involving minors). Pre-registration of hypotheses and methodologies are scientific best practices ensuring that (i) researchers don't change the rules halfway through to get the results they want; (ii) The employed process is transparent and replicable, so others can check the work; (iii) This reduces bias by committing to a plan before seeing any results. As a common practice in high-quality scientific studies, pre-registration promotes honesty, objectivity, and accountability, especially when the findings may be used to inform policy or public decisions.

experiments must be run in parallel and repeated over time to account for drift, retraining cycles, or time-of-day effects.

235. ***Comprehensive Logging***: Every piece of recommended content should be captured with timestamped metadata including its position in the feed, the algorithmic pathway (e.g., For You vs. Search), source account characteristics, and user engagement logs. This dataset should then be made available for auditing and replication purposes.

236. ***Blinded Annotation***: To evaluate exposure, all logged content should be labeled by multiple blinded raters using a predefined codebook. Reliability statistics (e.g., Cohen's κ or Krippendorff's α) should then be reported to ensure consistency.

237. ***Statistical Analysis***: Analysis must use appropriate hypothesis testing and effect size estimation. For binary outcomes (e.g., exposure to harmful content), chi-squared tests or logistic regression may be appropriate. For prevalence comparisons, z-tests for proportions or bootstrapped confidence intervals can quantify uncertainty. Causal inference techniques such as difference-in-differences (DiD) or matched group comparisons may also be employed when longitudinal or quasi-experimental designs are available.

238. By adhering to these principles, researchers can credibly isolate causal effects, quantify the reliability and scale of potential exposure harms, and ensure that their methodology is both scientifically transparent and statistically valid.

2. Mr. Bejar's Algorithmic Stress Test Suffers from Fatal Methodological Flaws Rendering the Results Scientifically Unreliable

239. Mr. Bejar's methodology suffers from the following core deficiencies:

240. ***Lack of Experimental Design***: There is no evidence that Mr. Bejar employed an experimental design worthy of the name. A valid experimental design necessarily includes a predefined hypothesis, experimental or control groups, randomized treatments, and standardized account initialization protocols. Mr. Bejar's test lacks evidence of each of these features. His experimental approach appears arbitrary in critical respects. He conceived of the idea independently after reading an article in the Wall Street Journal about the prevalence of certain

content on Instagram,[347] but he does not appear to have consulted any sources or collaborated with any qualified experts to aid in the design or direction of the experiment.[348] Further, there is no indication Mr. Bejar developed a set of specific criteria for the type of content recommendations he sought to assess, identified the relevant user populations, or otherwise set any particularized, time-limited goals for the test. On the contrary, Mr. Bejar took a casual approach guided by convenience and his personal whims. Mr. Bejar created "approximately seven" synthetic accounts as minor males and females[349] and conducted "initial testing" (which lasted for several minutes at most for each account) to see whether Reels would recommend "racy" or "violent" content.[350] But at some point he stopped checking most of the accounts after initial testing because he found the search results emotionally distressing.[351] During this pause, without methodological explanation, Mr. Bejar occasionally searched the phrase "I want to hurt myself" on one or more of the accounts,[352] and may have resumed other searches, but would stop soon again because he found the content disturbing.[353] When Mr. Bejar returned more fully to his project approximately a year after he started, he no longer checked the minor male accounts and instead focused on the minor female accounts because he "had all of them set up in a single phone and that made it easier."[354] In short, Mr. Bejar allowed his personal sensitivities and whims delineate his execution of the test. In academic literature, robust algorithmic audits require pre-registration or structured protocols that specify how synthetic accounts are configured, how engagement behaviors are scripted, and what constitutes a treatment versus a control condition. Mr. Bejar provides no such evidence or

[347] Arturo Bejar Dep. Vol. 3 at 995:13-25, Apr. 9, 2025.

[348] Arturo Bejar Dep. Vol. 3 at 1008-09, Apr. 9, 2025.

[349] Arturo Bejar Dep. Vol. 3 at 994:8-9, Apr. 9, 2025.

[350] Arturo Bejar Dep. Vol. 1 at 199:8-15, Apr. 7, 2025 (explaining that initial testing lasted anywhere from "8 to 12 minutes, sometimes 18").

[351] Arturo Bejar Dep. Vol. 3 at 998:8-11, Apr. 9, 2025.

[352] Arturo Bejar Dep. Vol. 3 at 998:18-21, Apr. 9, 2025

[353] Arturo Bejar Dep. Vol. 3 at 999:5-8, Apr. 9, 2025

[354] Arturo Bejar Dep. Vol. 3 at 999:1-8, 997:12-17, Apr. 9, 2025.

documentation. The absence of these elements renders the protocol scientifically uninterpretable and prevents any possibility of causal inference.

241. ***Confirmation and Interaction Bias***: The test accounts were clearly guided through intentional behaviors designed to surface borderline or policy-violating content. This includes searching for sensitive hashtags and engaging with content flagged as suggestive. For example, on one test account, Mr. Behar expressly "searched for some gymnastics content" which eventually led him to "videos of very young girls in different kinds of outfits."[355] On other accounts, Mr. Bejar searched the phrase "I want to hurt myself" an indeterminate number of times over a yearlong period.[356] On other test accounts, Mr. Bejar claims he did not conduct any searches, but quickly swiped through Reels videos without finishing until he came across content he deemed harmful (e.g., "a little violent" or "a little racy").[357] If the user posting the video appeared to Mr. Bejar to be under the age of thirteen, Mr. Bejar would watch the video to completion.[358] These forms of active user engagement are known to influence recommender systems,[359] which are designed to tailor content based on user behavior. This creates a feedback loop, [360] known in the literature as *interaction bias,* which reinforces the types of content a user engages with. Without neutral baselines or randomized interaction behaviors, the experiment cannot distinguish between service-driven recommendations and user-driven discovery.[361]

242. ***Opacity and Poor Documentation***: The documentation accompanying Mr. Bejar's

---

[355] Arturo Bejar Dep. Vol. 1 at 198:20-25, Apr. 7, 2025.

[356] Arturo Bejar Dep. Vol. 2 at 402:21-24, Apr. 8, 2025.

[357] Arturo Bejar Dep. Vol. 1 at 199:1-15, Apr. 7, 2025.

[358] Arturo Bejar Dep. Vol. 1 at 199:10-12, Apr. 7, 2025

[359] Baeza-Yates, R. (2020, September). Bias in search and recommender systems. In Proceedings of the 14th ACM conference on recommender systems (pp. 2-2).

[360] Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., & Burke, R. (2020, October). *Feedback loop and bias amplification in recommender systems*. In Proceedings of the 29th ACM international conference on information & knowledge management (pp. 2145-2148).

[361] Khenissi, S., & Nasraoui, O. (2020). Modeling and counteracting exposure bias in recommender systems. arXiv preprint arXiv:2001.04832.

materials consists primarily of editorialized file names, screen recordings, and fragmented screen captures. There is no formalized protocol outlining how many accounts would be used, over what time period the testing would be conducted, whether any efforts would be made to replicate findings, or how the outputs would be categorized and labeled. The only written record of the test consists of the so-called "test protocol" that Mr. Bejar drafted in 2025, long after the start of Mr. Bejar's inquiry in 2023.[362] The fact that the protocol was drafted long after the start of testing provides little confidence that the results of Mr. Bejar's efforts followed from the guidance the protocol sets forth. Moreover, Mr. Bejar's protocol is high-level, conclusory, and lacks all of the details outlined above that would even permit drawing reliable connections between the protocol and the results. Without purporting to identify all examples, the following are several of the critical gaps between Mr. Bejar's claims and his drafted protocol. Most critically, the protocol offers no transparency around whether content was organically recommended by the service or was the result of user searches or manual navigation. Further, the protocol offers no support for Mr. Bejar's assertion that he recorded video from each testing session,[363] never mind his admission that in some cases he only recorded part of the testing session.[364] Further, the protocol makes no mention that Mr. Bejar paused testing nor that he abandoned his consideration of the minor male accounts upon resuming his experiment.[365] Nor does the protocol support Mr. Bejar's claim that he planned testing sessions in advance, nor that Mr. Bejar would erase the iPhone 11 and redownload Instagram between testing sessions for each synthetic account.[366] As another example, Mr. Bejar claims he used, at various times, an iPhone 11, and at other times an iPad, but the protocol mentions

[362] Arturo Bejar Dep. Vol. 3 at 1001:12-16, Apr. 9, 2025; *see* BEJAR0002658.

[363] Arturo Bejar Dep. Vol. 3 at 1020:9-22, Apr. 9, 2025.

[364] Arturo Bejar Dep. Vol. 3 at 1021:1-5, Apr. 9, 2025. For example, Mr. Bejar admitted that he did not consistently capture some of his testing sessions wherein he searched the phrase "I want to hurt myself" and similar phrases. *Id.* at 1027-28.

[365] Arturo Bejar Dep. Vol. 3 at 1005:1-18, Apr. 9, 2025.

[366] Arturo Bejar Dep. Vol. 3 at 1019, Apr. 9, 2025.

only the former.[367] This lack of documentation makes the protocol entirely non-replicable and non-auditable.

243. ***No Statistical Rigor or Validity***: Mr. Bejar provides no statistical measurement of exposure rates, prevalence of harmful content, latency of content surfacing, or control comparisons across different account behaviors. He does not define any metrics for measuring algorithmic performance (e.g., precision, recall, rate of harmful content exposure), nor does he employ any tools for validating whether the content exposure observed was anomalous or consistent with typical service behavior. The sample size is undefined or too small to yield a statistically appropriately powered study; no effort is made to perform statistical analysis across multiple trials. The results that are shown are the byproduct of "statistical cherry-picking",[368] i.e., the malpractice of selecting evidence that corroborates a position, while discarding evidence of the contrary: Mr Bejar excluded the results from a subset of synthetic accounts he created, approximately half of the total, without any explanation or justification. These deficiencies disqualify the methodology from being considered scientifically valid.

244. ***Failure to Disentangle Algorithmic Output from User Behavior***: At its core, Mr. Bejar's approach confounds the effects of service recommendation mechanisms with user steering behavior. [369] Many of the outputs catalogued appear to follow naturally from the actions taken by the user (e.g., lingering on, liking, or replaying suggestive videos while ignoring any other type of content). Without controlling for these behaviors, or at least holding them constant across experimental conditions, it is impossible to determine whether the recommender system is surfacing harmful content organically or simply responding to input behavior, as designed. No

[367] Arturo Bejar Dep. Vol. 3 at 1011-13, Apr. 9, 2025.

[368] *Cherry picking*. (n.d.). Wikipedia. Retrieved June 13, 2025, from https://en.wikipedia.org/wiki/Cherry_picking#:~:text=Cherry%20picking%2C%20suppressing%20evidence%2C%20or,that%20may%20contradict%20that%20position.

[369] Chaney, A. J., Stewart, B. M., & Engelhardt, B. E. (2018, September). *How algorithmic confounding in recommendation systems increases homogeneity and decreases utility*. In Proceedings of the 12th ACM conference on recommender systems (pp. 224-232).

counterfactuals are presented to test alternate explanations.

245. Accordingly, Mr. Bejar's findings cannot be considered reliable or generalizable assessments of how Instagram's recommendation systems function. His results are better characterized as anecdotal observations lacking the scientific safeguards required to support meaningful conclusions. Simply put, what Mr. Bejar conducted is not a "stress test" in any scientifically accepted sense; it is a series of anecdotal observations devoid of methodological grounding.

246. Mr. Bejar's findings do not provide meaningful evidence of how Instagram's recommender systems function either under normal circumstances or under simulated stress conditions. They do not demonstrate systemic failure in content moderation, algorithmic bias, or harmful amplification. They provide no insight into algorithmic thresholds, classifier behavior, or enforcement mechanisms. In short, Mr. Bejar's materials are not only methodologically unsound, but they are also fundamentally unreliable and should carry no evidentiary weight in any serious inquiry into service integrity.

VII. Meta's Content Moderation Policies & Enforcement Align with Industry Best Practices

247. Meta's content moderation policies, the technology and review processes that enforce them, as well as the public reporting of those efforts align with best practices in the technology industry.

248. The Digital Trust & Safety Partnership ("DTSP") has developed a widely accepted framework that articulates industry best practices for managing content- and conduct-related risks across five key commitments: (i) risk evaluation in product development, (ii) product governance, (iii) enforcement operations, (iv) iterative improvement, and (v) transparency.[370] In my expert opinion, Meta's content moderation infrastructure and governance systems not only align with the

[370] Digital Trust & Safety Partnership. (n.d.). *Trust & Safety Best Practices Framework*. Retrieved June 11, 2025, from https://dtspartnership.org/wp-content/uploads/2021/04/DTSP_Best_Practices.pdf.

DTSP's framework but in many areas serve as a leading implementation of these best practices.

A. Product Development: Proactive Risk Identification and Mitigation

249. Meta integrates Trust & Safety considerations from the earliest stages of product development. As detailed in this report, Meta convenes biweekly Product Policy Forums that include safety and cybersecurity policy teams, product managers, and legal and policy leads to assess risk before feature deployment. The company's content policy team, informed by internal analyses and external consultations (e.g., with mental health and child safety organizations), participates directly in product iteration. This mirrors DTSP's recommendation that companies embed risk assessment, stakeholder input, and feedback loops into the development lifecycle.

B. Product Governance: Transparent and Evolving Policy Structures

250. Meta maintains a layered governance structure consisting of its public Community Standards, internal Policy Labs, and Borderline Content Policies. These are developed with input from academics, civil society groups, and domain experts. Changes to policy are logged and made publicly visible. These practices reflect DTSP's emphasis on explainable rulemaking, user-facing clarity, community input, and documented internal interpretative standards.

C. Enforcement Operations: Scalable and Responsive Moderation Infrastructure

251. Meta operationalizes product governance through a hybrid enforcement system involving (i) scalable AI classifiers and proactive detection, (ii) global human review teams including specialized "ring-fenced" experts, (iii) tiered prioritization mechanisms such as Predictive Integrity Value (pIV) and the HERO system, and (iv) robust user reporting and appeals systems. Meta also provides employee wellness support and workload safeguards for reviewers, directly aligning with DTSP guidance on workforce resilience, reporting tools, and scaled enforcement workflows.

D. Iterative Improvement: Data-Driven Learning and Policy Adjustment

252. Meta continuously updates its classifiers, policy enforcement thresholds, and risk mitigation strategies based on feedback from audits, real-world incidents, and statistical performance reviews. The use of systems like Label Accuracy Management (LAM), the

enforcement metrics reported in CSERs (e.g., proactive detection rates), and ongoing adjustments based on expert consultation exemplify DTSP's call for structured learning loops and explainable frameworks for policy evolution.

E. Transparency and Accountability: Industry-Leading Disclosures

253. Meta publishes detailed Community Standards Enforcement Reports (CSERs) quarterly, providing metrics on prevalence, proactive rate, appeals, and restoration. These reports include confidence intervals, describe classifier methodology, and enable public scrutiny. Meta also engages with researchers through dataset access initiatives, sponsors benchmarking challenges (e.g., Deepfake Detection and Hateful Memes), and incorporates in-product enforcement signals such as warning screens and user notices. These practices are consistent with DTSP's call for public-facing policy disclosures, support for academic collaboration, and periodic reporting to stakeholders.

254. In sum, Meta not only fulfills the commitments described in the DTSP Best Practices Framework but does so at a scale and level of transparency that sets a benchmark for the industry. Based on my review of Meta's content moderation infrastructure, I conclude that Meta is a paradigmatic example of a DTSP "Practicing Company."

VIII. Conclusions

255. Based on my review of the evidence, internal documentation, publicly disclosed data, and relevant technical and scientific literature, I offer the following expert findings regarding Meta's content moderation systems:

256. Meta has implemented one of the most technically advanced and operationally mature content moderation infrastructures in the industry. Its multi-pronged approach includes AI-based classifiers and other automated detection systems, human review pipelines, escalation procedures, and reactive user-report channels, all of which are integrated into a coordinated enforcement framework.

257. Meta's Community Standards Enforcement Reports ("CSERs") provide a scientifically sound and statistically robust mechanism for evaluating the effectiveness of content

moderation. The "prevalence" metric, in particular, offers a meaningful proxy for potential user harm by measuring actual exposure to violating content.

258. Meta's moderation systems prioritize high-severity harms, such as suicide, self-injury, and child exploitation, and demonstrate exceptionally high proactive detection rates in these categories, often exceeding 98%. These rates reflect the maturity and effectiveness of Meta's AI classifiers, as well as the prioritization mechanisms embedded in its enforcement architecture.

259. The company has demonstrated a transparent and iterative approach to measuring, reporting, and refining its moderation systems. The inclusion of appeal and restoration metrics, alongside published methodologies, underscores a commitment to accountability and continuous improvement.

260. Meta's systems are not static but are designed to evolve in response to emerging harms, adversarial tactics, cultural considerations, and regulatory constraints. Classifier thresholds, enforcement priorities, and borderline content protocols are continuously updated based on new evidence and service dynamics.

261. Allegations that Meta's recommendation systems promote harmful content, including those based on anecdotal stress tests such as the one conducted by Mr. Bejar, are not grounded in scientifically valid methodologies. Such audits fail to meet basic requirements for control, replication, statistical rigor, or causal inference, and should not be afforded evidentiary weight in serious evaluations of service integrity.

262. Taken together, the evidence supports the conclusion that Meta's content moderation systems are robust, reasonably designed, and continuously improving in ways that are consistent with both academic best practices and industry standards.

Emilio Ferrara, Ph.D.

June 13, 2025

Appendix A: Table of Harm-Specific Classifiers

Plain-language classifier name	Alpha-numeric code	Launch date of current iteration	Threshold value for autodeletion	Threshold value for soft actions
Suicide & Self-Injury				

SSI Classifier	F590828509	08/12/2024	.94	Facebook: various enforcements >P50
				• EU Demotion on Newsfeed
				  ○ Suicide: 0.8
				  ○ Self-injury: 0.8
				• EU Deboost on Newsfeed
				  ○ Suicide: 0.3
				  ○ Self-injury: 0.4
				• Non-EU Demotion on Newsfeed
				  ○ Suicide: 0.5
				  ○ Self-injury: 0.5
				• Non-EU Deboost on Newsfeed
				  ○ Suicide: 0.3
				  ○ Self-injury: 0.4
				• Age-based Demotion on Newsfeed
				  ○ Suicide: 0.3
				  ○ Self-injury: 0.3
				• Age-based Deboost on Newsfeed
				  ○ Suicide: 0.05
				  ○ Self-injury: 0.05
				• Age gating on Newsfeed
				  ○ Suicide: 0.1
				  ○ Self-injury: 0.1
				• Non-rec filtering
				  ○ Suicide: 0.105
				  ○ Self-injury: 0.25
				• Filtering demotion on Stories
				  ○ Self-injury: 0.25
				• Age gating on Stories
				  ○ Self-injury: 0.25
				• Age gating on Stories (reduce more)
				  ○ Self-injury: 0.1

				Instagram: various enforcements > P40
				• Age gating
				  ○ Suicide: 0.89
				  ○ Self-injury: 0.71
				• Age gating High COE
				  ○ Suicide: 0.6
				  ○ Self-injury: 0.24
				• Non-rec filtering
				  ○ Suicide: 0.6
				  ○ Self-injury: 0.24
				• Hashtag filtering
				  ○ Suicide: 0.80
				  ○ Self-injury: 0.80
				• Media SERP
				  ○ Suicide: 0.85
				  ○ Self-injury: 0.85
				• Age-filtering
				  ○ Suicide: 0.6
				  ○ Self-injury: 0.24
				• Demotion Feed teen non-EU
				  ○ Suicide: 0.7
				  ○ Self-injury: 0.7
				• Demotion Feed teen EU
				  ○ Suicide: 0.85
				  ○ Self-injury: 0.85
				• Demotion Stories teen
				  ○ Suicide: 0.7Self-injury: 0.7

SSI Borderline Classifier	f551409441	04/15/2024	N/A	Facebook: various enforcements above >P10 • Age gating on Newsfeed: 0.1 • P-non MSI: 0.3 • Non-rec filtering: 0.3 • Age gating on Stories: 0.75 • Search demotion: 0.5 Instagram: various enforcements > P40 • Age gating: 0.91 • Age gating High CoE: 0.54 • Non-rec filtering: 0.34 • Media SERP: 0.85

				• Hashtag filtering: 0.80 • Age filtering: 0.25
Eating Disorder				
ED Classifier	f638461582	09/01/2024	N/A	Facebook • EU Demotion on Newsfeed: 0.8 • EU Deboost on Newsfeed: 0.15 • Non-EU Demotion on Newsfeed: 0.5 • Non-EU Deboost on Newsfeed: 0.15 • Age based Demotion on Newsfeed: 0.3 • Age based Deboost on Newsfeed: 0.05 • Age gating on Newsfeed: 0.1 • Non-rec filtering: 0.061 Instagram: various enforcements >P40 • Non-rec filtering: 0.81 • Media SERP: 0.85 • Hashtag filtering: 0.80 • Demotion Feed teen non-EU: 0.7 • Demotion Feed teen EU: 0.85 • Demotion Stories teen: 0.7
ED Borderline Classifier	f551542423	04/15/2024	N/A	Facebook • Age gating on Newsfeed: 0.1 • P-non MSI: 0.3 • Non-rec filtering: 0.33 • Search demotion: 0.5 Instagram: various enforcements >P40 • Age gating: 0.907 • Non-rec filtering: 0.7 (P60) • Media SERP: 0.85 • Hashtag filtering: 0.80 Age filtering: 0.26 (P40)
Bullying & Harassment				

FB Comment	8165225660248547	10/20/2024	Bullying and Harassment = 0.95 Borderline Hostile Speech = n/a	Bullying and Harassment: • Filtering on Comments: 0.432 • Age filtering comments: 0.275 Borderline Hostile Speech: • Filtering on Comments: 0.6 • Age filtering on Comments: 0.2
FB Post	8655505921181700	10/23/2024	Bullying and Harassment = 0.95 Borderline Hostile Speech = n/a	Bullying and Harassment: • Age-based Non-rec Filter on Newsfeed: P30 • Age-based Deboost on Newsfeed: P10 • Age-based Demotion on Newsfeed: P30 • Deboost on Newsfeed: P50 • Demotion on Newsfeed: P50 • EU Demotions on Newsfeed: P80 • Demotion on IFR: 0.50 • Non-rec Filter on IFR: 0.60 • Age gating on Newsfeed: P10 • ARC Demotion on Newsfeed: P50 Borderline: • Age-based Non-rec Filter on Newsfeed: P10 • Demotion on IFR: 0.50 • Non-rec Filter on IFR: P60 • ARC Demotion on Newsfeed: P50 • ARC Deboost on Newsfeed: P20
IG Comment	5295222083900364	06/06/2022	Bullying and Harassment = 0.95	Bullying and Harassment: • Comment filter (auto-hide): P80 • Comment cover: P70 • Advanced comment filtering: P60

			Borderline Hostile Speech = n/a	• Comment downranking (demotion): P50 • Preview comment filter: P50 Borderline Hostile Speech: • Comment filter (auto-hide): P95 • Comment cover: P80 • Advanced comment filtering: P70 • Comment downranking (demotion): P70 • Preview comment filter: P70
IG Post	24098062513172542	11/05/2023	Bullying and Harassment = 0.95 Borderline Hostile Speech = n/a	Bullying and Harassment: • Youth non-rec filter: P50 • Youth demotions on connected stories: P60 • Youth demotions on connected feed: P60 • EU youth demotions on connected feed: P80 Borderline Hostile Speech: • Youth non-rec filter: P40
Reels	7408044142611719	04/27/2024	Bullying and Harassment = 0.95 Borderline Hostile Speech = n/a	Facebook Bullying and Harassment: • Age-based Non-rec Filter on Reels: P15 • Non-rec Filter on Reels: P25 • Non-rec Filter on Watch: P25 Borderline Hostile Speech: • Age-based Non-rec Filter on Reels: P15 • Non-rec Filter on Reels: P25 • Non-rec Filter on Watch: P25 Instagram Bullying and Harassment: • Youth demotions on connected feed: P60

				• EU youth demotions on connected feed: P80 Borderline Hostile Speech: • Youth non-rec filter: P40
Graphic and Violent Content				

Graphic Violence MAD photo classifier on FB	f647163274	10/01/2024	N/A	• Mark as disturbing ("MAD"): 0.972
				• Various enforcements > P50
				  ○ Reels: 0.274
				  ○ CFR Demotion: 0.274
				  ○ CFR Deboost: 0.274
				  ○ Stories demotion: 0.2529
				  ○ Search filtering : 0.2529

Graphic Violence MAD video classifier on FB	f645517662	9/25/2024	N/A	• MAD: 0.98
				• Various enforcements > P50
				  ○ Reels 0.2529
				  ○ CFR Demotion: 0.2529
				  ○ CFR Deboost: 0.2529
				  ○ Stories demotion: 0.2529
				  ○ Search filtering : 0.2529

Graphic Violence MAD photo classifier on IG	f647151014	09/30/2024	N/A	• MAD: 0.967 • EU age gating on Feed: 0.8 • Non-EU age gating on Feed: 0.6
Graphic Violence MAD video classifier on IG	f639562790	09/11/2024	N/A	• MAD: 0.91 • EU age gating on Feed: 0.8 • Non-EU age gating on Feed: 0.6
GV Video Deletion classifier on IG	3614043688651385	02/21/2021	0.62	N/A
Borderline GV classifier			N/A	Facebook Youth: Various enforcements > P10 • Feed age gating: 0.03 (P10) • Feed demotion: 0.12 (P30) • Rees age gating: 0.06 (P20) • Reels age filtering: 0.09 (P25)

				• Watch/IFR filtering: 0.09 (P25) Instagram Youth: various enforcements > P40 • Filtering: 0.341 (P50) • Demotion: 0.204 (P40) • Age gating: 0.653 (P80)
Adult Nudity & Sexual Activity				
Violating ANSA photo classifier on FB and IG	f644294876	09/18/2024	• 0.952 for FB • 0.96 for IG	Instagram • Age gating: 0.8 Facebook Youth filter on stories: 0.5 Youth filter on reels: 0.275 Youth filter on watch: 0.5 Youth filter on feed: 0.1
Violating ANSA video classifier on FB and IG	f643324421	09/16/2024	• 0.981 for FB 0.993 for IG	Instagram • Age gating: 0.8 Facebook • Youth filter on stories: 0.5 • Youth filter on reels: 0.275 • Youth filter on watch: 0.5 Youth filter on feed: 0.1
Violating ANSA linkshare classifier on FB and IG	f635467640	08/26/2024	0.55 for FB	Instagram • Age gating: 0.8 Facebook • Youth filter on stories: 0.5 • Youth filter on reels: 0.275 • Youth filter on watch: 0.5 • Youth filter on feed: 0.1
Borderline ANSA Photo Classifier on FB	f645114003	9/21/2024	N/A	• Youth filter on Reels: 0.055 • Youth filter on Watch: 0.095 • Youth filter on Feed: 0.15 • Youth filter on IFR: 0.1 • Youth filter on Story: 0.5
Borderline ANSA Video Classifier on FB	f645120189	9/21/2024	N/A	• Youth filter on Reels: 0.055 • Youth filter on Watch: 0.095 • Youth filter on Feed: 0.15 • Youth filter on IFR: 0.1 • Youth filter on Story: 0.5

IG Non-rec ANSA Photo Classifier on IG	f639524943	09/04/2024	N/A	• Age gating: 0.623 • Age filtering: 0.159 • Age demotion: 0.104
IG Non-rec ANSA video classifier on IG	f641128648	09/10/2024	N/A	• Age gating: 0.623 • Age filtering: 0.237 • Age demotion: 0.158
Child Safety				
CSAM classifier	FB: f588692710; IG: f565741344	08/29/2024	N/A	Facebook • Age filtering: P20 • Filtering for GenPop: P25 Instagram Filtering for Teens: P25
CSAM Solicitation classifier	FB: f636246456; IG: f566561739	09/16/2024	N/A	Facebook • Age filtering: P20 • Filtering for GenPop: P25 Instagram Filtering for Teens: P25
CSAM-S comment classifier	f636721225	08/29/2024	N/A	Facebook • Age filtering: P20 • Filtering for GenPop: P25 Instagram IG: no soft actions
Child Sexualization (CSx) classifier	f590979030	08/28/2024	N/A	Facebook • Age filtering: P20 • Filtering for GenPop: P25 Instagram Filtering for Teens: P25
CSE-I Classifier	f573679609	06/18/2024	N/A	Facebook • Age filtering: P20 • Filtering for GenPop: P25 Instagram IG: no soft actions
Child subject to objectionable content classifier	f589982602	08/20/2024	N/A	Facebook • Age filtering: P20 • Filtering for GenPop: P30 Instagram Filtering for Teens: P50
Google Content Safety API classifier	N/A	06/02/20	N/A	No soft actions